VOLUME V JANUARY, 1926 NUMBER 1

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

50c per copy *$1.50 per year*

THE BELL SYSTEM TECHNICAL JOURNAL

Published quarterly by the American Telephone and Telegraph Company, through its Information Department, in behalf of the Western Electric Company, the Bell Telephone Laboratories, Inc., and the Associated Companies of the Bell System

Address all correspondence to the Editor
Information Department
AMERICAN TELEPHONE AND TELEGRAPH COMPANY
195 BROADWAY, NEW YORK, N. Y.

50c. Per Copy

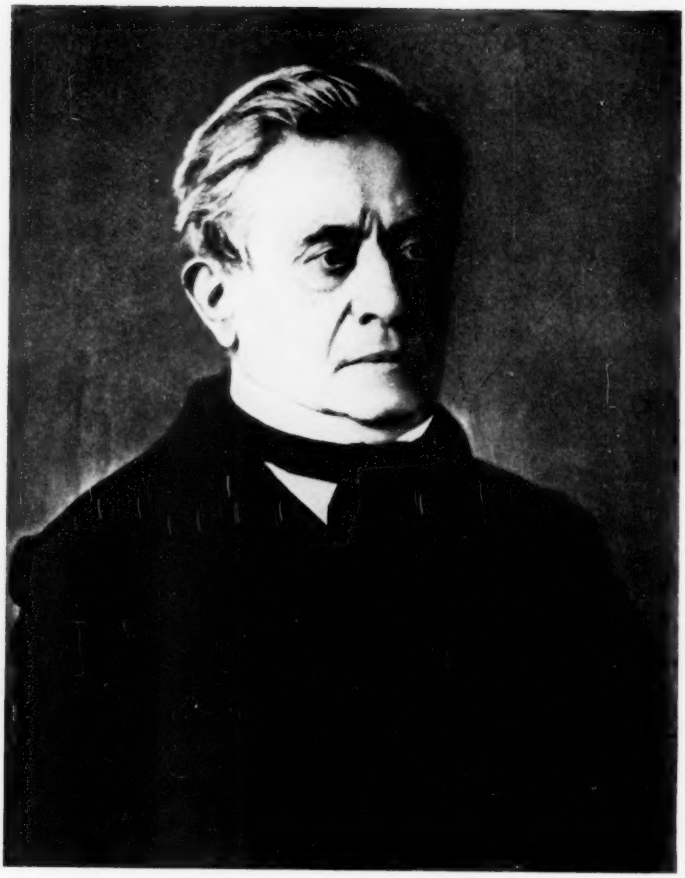
$1.50 Per Year

EXPLANATORY

Previous to 1907 there were maintained in the Bell System three laboratories and departments of development, research and experiment,—one by the American Telephone and Telegraph Company at Boston, one by the Western Electric Company at Chicago and one by the Western Electric Company at New York. In 1907, in the interest of economy and efficiency, these were consolidated so far as laboratory and experimental work were concerned and the Bell System Laboratory at Bethune and West Streets, New York, was established. This was incorporated as the Bell Telephone Laboratories, Inc., January first, 1925. The expense of operation is divided between the American Telephone and Telegraph Company and the Western Electric Company according to the nature of the work done.

In the Bell System the American Telephone and Telegraph Company undertakes, through constant association with the operating organizations, to formulate the requirements, present and future, of the Bell System. Out of these requirements come the problems of the American Telephone and Telegraph Company's Department of Development and Research and the System laboratory. After the problems have been satisfactorily solved, the Department of Development and Research adopts as standard the systems, equipment and apparatus thus produced which are then specified for their proper uses in the Associated Companies by the Engineering Department of the American Telephone and Telegraph Company. When different departments and companies are mentioned in this publication, so far as they are parts of the Bell System, they are parts of *one* working organization.

H. B. THAYER, *Chairman*,
American Telephone and Telegraph Company.



JOSEPH HENRY

1799–1878

The Bell System Technical Journal

January, 1926

Joseph Henry

The American Pioneer in Electrical Communication

By BANCROFT GHERARDI and ROBERT W. KING

IN the brilliant galaxy of investigators to whom we owe our knowledge of electrical science, Joseph Henry stands out as of the first magnitude; and for those who are associated with the Bell System, the present is a most appropriate time to review his researches which had an important guiding influence on the development of electrical communication. The present year marks the fiftieth since the invention of the telephone by Alexander Graham Bell, and among the scientists with whom Bell conferred at that time, he gave a place of honor to Henry. In a letter to his parents written in March, 1875, while he was busy in an effort to perfect the harmonic telegraph, and before he had turned his attention to the telephone, Bell wrote:

"Now to resume telegraphy. When I was in Washington, I had a letter of introduction to Professor Henry, who is the Tyndall of America. I had found on inquiry at the Institute of Technology, that some of the points I had discovered in relation to the application of acoustics to telegraphy had been previously discovered by him. I thought I would, therefore, explain all the experiments, and ascertain what was new and what was old. He listened with an unmoved countenance, but with evident interest to all, but when I related an experiment that at first sight seems unimportant, I was startled at the sudden interest manifested.

"I told him that on passing an intermittent current of electricity through an empty helix of insulated copper wire, a noise could be heard proceeding from the coil, similar to that heard from the telephone. He started up, said, 'Is that so? Will you allow me, Mr. Bell, to repeat your experiments, and publish them to the world through the Smithsonian Institute, of course, giving you the credit of the discoveries?'

"I said it would give me extreme pleasure, and added that I had apparatus in Washington, and could show him the experiments myself at any time. . . .

"We appointed noon next day for the experiments. I set the in-

strument working and he sat at a table for a long time with the empty coil of wire against his ear listening to the sound. I felt so much encouraged by his interest that I determined to ask his advice about the apparatus I have designed for the transmission of the human voice by telegraph. I explained the idea and said, 'What would you advise me to do, publish it and let others work it out, or attempt to solve the problem myself?' He said he thought it was 'the germ of a great invention,' and advised me to work at it myself instead of publishing. I said that I recognized the fact that there were mechanical difficulties in the way that rendered the plan impracticable at the present time. I added that I felt that I had not the electrical knowledge necessary to overcome the difficulties. His laconic answer was, 'GET IT.'

"I cannot tell you how much these two words have encouraged me. Such a chimerical idea as telegraphing vocal sounds would indeed to most minds seem scarcely feasible enough to spend time in working over. I believe, however, that it is feasible, and that I have got the cue to the solution of the problem.

"Professor Henry seemed to be much interested in what I told him, and cross-questioned me about my past life, and specially wanted to know where I had studied physics . . . . "

Joseph Henry was born in Albany, New York, in 1799, and coming to full maturity of mind at the beginning of a century which will probably never be surpassed for fruitful research in the field of electricity, he demonstrated, at the very outset of his career, his right to stand for all time with the foremost investigators in this department of natural science. Henry was, moreover, a many-sided man. His distinguished career leads into many fields and before reviewing his researches on electro-magnetism we may note briefly the very diversified and yet important character of his other work.

During the latter half of his life, official duties as the director of the Smithsonian Institution consumed an ever increasing portion of his time, but he still found opportunity to prosecute many original inquiries,—for example, into the application of acoustics to building, into the best construction and arrangement of lecture rooms, and into the strength of various building materials. As one of his first administrative acts, he organized a widespread corps of observers for simultaneous weather and meteorological reports by means of the telegraph which was yet in its infancy. He was the first to have the daily atmospheric conditions indicated upon a map of the country and to utilize this information in making weather forecasts.

He was an active and long-standing member of the Lighthouse Board of this country and his diligent investigations into the efficiency of various illuminants and the best conditions for their use greatly improved the beacons which dotted our coasts. During the dark days of the Civil War, Henry clearly saw the tremendous advantage to be derived from a mobilization of the nation's scientific men for cooperative service. His vision, backed by his tremendous energy and ability, resulted in the formation of the National Academy of Sciences, under a Congressional charter signed by Abraham Lincoln.

More than fifty years later this same National Academy of Sciences was again called upon in time of national need, and, using the mechan-

Fig. 1

ism inspired by Henry, there re-occurred, in 1916, under presidential proclamation, a mobilization of the nation's scientific and technical men.

While the details of Henry's life and work are perhaps not widely

known, his researches are of the most enduring character, and for all time must enter intimately into the lives of all civilized mankind. He was without peer among the American physicists of his time, and it is well attested by every record that he was a man of varied culture, of large breadth and liberality of views, of generous impulses, of great gentleness and courtesy of manner, combined with equal firmness of purpose and energy of action.

Let us now turn to Henry's investigations of electro-magnetism, which were among his earliest scientific undertakings. He began his career in 1826 in New York State at the Albany Academy, where he had only the apparatus he could construct with his own hands and, out of each year, but a single month uninterrupted by other duties to devote to his researches. It was there—independently of Faraday and on some fundamental points prior to him—that Henry discovered the laws of current induction. At the same time he undertook a study of the electromagnet which prepared the way for not only the telephone and telegraph, but also for all types of dynamos and motors.

The electromagnet was discovered by Sturgeon in England, but Henry's contributions to our knowledge of it were so great that after his work, a powerful instrument suitable for many uses replaced what had been a feeble toy. When he started his work on the electromagnet its design was not understood; when he had completed his work he had developed a magnet, the design of which was understood and which could be adapted, according to the rules which he laid down, to a multitude of purposes.

With reference to the making of electromagnets, Henry pointed out the improvements which resulted from insulating the conducting wire itself, instead of the rod to be magnetized, and by covering the whole surface of the iron with a series of coils in close contact. This was effected by insulating a long wire with silk thread, and winding this around the rod of iron in close coils from one end to the other. The same principle was extended by employing a still longer insulated wire, and winding several strata of this over the first, care being taken to insure the insulation between each stratum by a covering of silk ribbon. By this arrangement the rod was surrounded by a compound helix formed of a long wire of many turns instead of a single helix of a few turns.

Thus Henry laid down the rules, which, in general, are followed today in the construction of commercial electromagnets; namely, that the wire should be insulated, that it should be wound in layers, and that there should be several layers, one above the other. He

also did another thing in his actual construction: he adopted what may be called the spool construction, the placing of the windings on spools, and then the sliding of the spools on the core. That is a standard method of building electromagnets today.

Soon after doing this work Henry built a magnet to be used at Yale University, which was in its time a wonder and would even today be considered very powerful. He also built a series of magnets in which the emphasis was placed upon the lifting power in relation to the weight of the magnet and succeeded in designing one which, when energized by a single small cell, could support 420 times its own weight.

The improvements which Henry made in magnets suggested to him applications of magnetic attraction to the production of mechanical motion. He realized that electromagnets such as he built were easy to control, and believed that he could design a machine by which he could get power from an electric current and this at a time when the only source of current were primary batteries as the dynamo did not yet exist.

His electric motor was the first ever built to use electromagnets;[1] it was extremely simple consisting of an electromagnet supported at its center by a pivot so that it could rock back and forth under the alternating pulls of two permanent magnets. To effect the reversal of magnetization of the electromagnet and hence the alternation of pulls, mercury cups were arranged so that wires would dip in them as the suspended magnet rocked to and fro. These contacts were the prototype of the commutator which is found in every direct current motor and dynamo today. It is interesting to note the words in which Henry described this invention. In Silliman's *American Journal of Science* for 1831 he wrote, "I have lately succeeded in producing motion in a little machine by a power which I believe has never before been applied in mechanics—by magnetic attraction and repulsion. Not much importance, however, is attached to the invention since the article in its present state can only be considered a philosophical toy; although in the progress of discovery and invention it is not impossible that the principle or some modification of it on a more extended scale may hereafter be applied to some useful purpose."

The modesty of this statement and Henry's vision of the future possible applications of the principle there shown cannot fail to com-

[1] Faraday has some years before shown that a wire carrying a current could be caused to revolve continuously around the pole of a permanent magnet. Henry's advance over this was considerable in that he materially increased the force causing motion by employing the attraction between two magnets, one permanent and one generated by current. The motor using electromagnets throughout did not come until later.

mand our admiration. Of course, until the dynamo was invented at a later date, and a substantial electric current became available, the motor could not be much more than he characterized it, "a philosophical toy."

Henry also became interested in a determining whether an electromagnet could be operated from a distance so that the doing of some work—for example the ringing of a bell—could be controlled from a distant station. From his investigations directed to this end, Henry was the first to appreciate that the effect of the resistance of long lengths of wire to the passage of electric current could be minimized by properly proportioning the battery and the magnet windings to the length and resistance of the line wires.

Efforts had been made by others prior to Henry's time to devise successful electric telegraphs. They had failed, however, because they did not know how to proportion their magnets and their batteries so as to operate over any substantial length of line. The literature of that time contains a number of demonstrations of the impossibility of operating an electric telegraph, because scientists could arrange instruments which would operate successfully when separated by a few feet, or even one hundred feet, but they would not work at a distance of thousands of feet because of the resistance of the long line wire.

What Henry did was to determine the proportioning of the various parts of the system so as to secure operation. He found, when his magnet was connected by a short wire to the battery, that the greatest magnetizing effect was obtained by joining the cells of the battery in parallel, but that a series arrangement of the battery would give the greatest pull if a long wire (a length of a mile or more was used in some of his experiments) carried the current. He also obtained the best operation over a short line when the magnet winding consisted of several distinct coils, all connected in multiple; and for operation over a long line he found it best either to connect these coils in series or to apply to the magnet a single long winding. Henry was therefore the first to produce an electric telegraph, and more than that, the transmission of electrical energy to a distance. That first telegraph paved the way for all the telegraph systems, all the ocean cable systems, and contained the principle of all telephone call bells.

One of Henry's greatest discoveries from the standpoint of electrical science, but a discovery in which he must yield the first place to Faraday, is that of mutual induction—the fact that a wire when moving with respect to a magnetic field has an electromotive force generated in it. Although Henry made his discovery independently

of Faraday, the latter was the first to make known his observations to the world, and it is no trifling index of Henry's character that he never in any way intimated that he was entitled to share with Faraday credit for the discovery.

Because Henry was anticipated in the publication of his observation of mutual induction, he does not appear to have left a verbal record of the steps of reasoning by which he was led to the discovery. However, he does tell us what the arrangement of apparatus was and if we bear in mind that he was seeking a method of generating an electric current from a magnet—this magnet, in turn, being itself the product

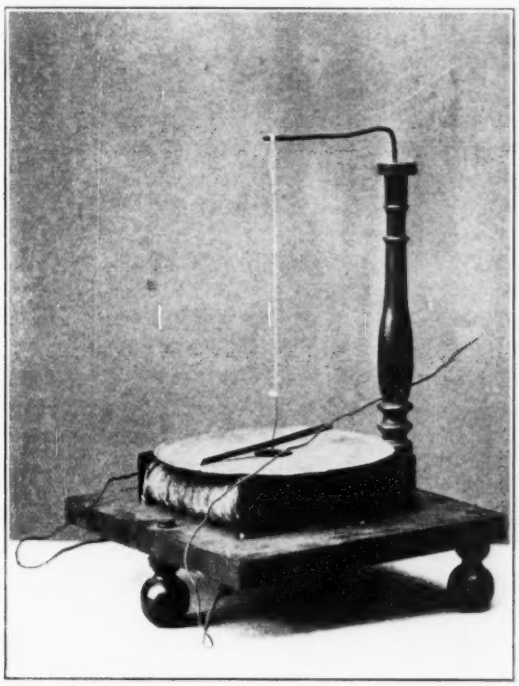


Fig. 2

of a current—we cannot but be impressed by the directness of his method.

Writing of his original observations, Henry says he "succeeded in producing electrical effects in the following manner, which differs

from that employed by Mr. Faraday and which appears to me to develop some new and interesting facts. A piece of copper wire, about thirty feet long and covered with elastic varnish, was closely coiled around the middle of the soft iron armature of a galvanic magnet . . . which, when excited will readily sustain between six and seven hundred pounds. The armature thus furnished with wire was placed in its proper position across the ends of the magnet and fastened so that no motion could take place. The two projecting ends of the helix were connected with a distant galvanometer by means of two copper wires each about forty feet long. This arrangement being completed, I stationed myself near the galvanometer and directed an assistant at a given word to suddenly immerse the galvanic battery attached to the magnet. At the instant of immersion the north end of the needle was deflected 30° to the west, indicating a current of electricity from the helix surrounding the armature. The effect, however, appeared only as a single impulse, for the needle after a few oscillations, resumed its former undisturbed position, although the action of the battery was still continued. I was, however, much surprised to see the needle suddenly deflected from a state of rest to about 20° to the east, when the battery was suddenly withdrawn from the acid, and again deflected to the west when it was re-immersed. This operation was repeated many times in succession, and uniformly with the same result."

It was in this same paper that Henry announced his observation of the phenomenon of self-induction, a most important discovery and one for which he holds full credit for having first made it known to the world. He writes, "I may, however, mention one fact which I have not seen noticed in any work, and which appears to me to belong to the same class of phenomena as those before described; it is this: when a small battery is moderately excited by diluted acid, and its poles, which should be terminated by cups of mercury, are connected by a copper wire not more than a foot in length, no spark is perceived when the connection is either formed or broken; but if a wire of thirty or forty feet long be used instead of the short wire, though no spark will be perceptible when the connection is made, yet when it is broken by drawing one end of the wire from its cup of mercury, a vivid spark is produced . . . . . The effect appears somewhat increased by coiling the wire into a helix." In a somewhat later paper we find the following statement. "A ribbon of sheet copper nearly an inch wide, and twenty-eight and a half feet long, was covered with silk, and rolled into a flat spiral similar to the form in which woolen binding is found in commerce. With this a

vivid spark was produced, accompanied by a loud snap. The same ribbon uncoiled gave a feeble spark."

Henry tried many modifications of this experiment and in the end drew the conclusion that the after-current he was observing was due to the inductive effect of the current in the wire upon itself, and that this became particularly apparent when the wire was so coiled that its various turns lay close together. The discovery of mutual induction by Faraday and the discovery of self-induction by Henry constitute two halves of a whole, and it is appropriate that to these men should go equal recognition in the matter of having electrical units names after them. Of the three units by which the properties of every electric circuit are measured, the unit of capacity was named after Faraday, and unit of inductance after Henry; the third unit, that of resistance, recognizes the fundamental researches of Ohm.

A few years later, after having accepted the chair of physics at Princeton University, Henry returned to the subject of induced currents. In his earlier work he, like Faraday, had used the continuous currents which a voltaic battery generates. He now chose the currents which flow when a Leyden jar is discharged. To register the inductive effects of the fleeting currents of discharge Henry adopted a device consisting of an unmagnetized needle placed in a small coil of wire. Through this coil the induced current had to flow. The use of the needle as an indicator led Henry to an important observation. He noticed that following a discharge, the direction of magnetization of the needle depended upon the distance across which the inductive effect had occurred. To account for this curious result, he advanced the hypothesis—later shown to be correct—that the discharge is oscillatory.

Here was the germ of a great discovery. The oscillatory character of the discharge is one of the fundamental and important properties of certain types of electric circuit. Henry did not have the facilities, however, for carrying his investigations in this field far enough to attract the attention of the scientific world. It was not until 1875, some thirteen years later, when Lord Kelvin was led independently by mathematical considerations to believe that the discharge is oscillatory, that the significance of the phenomenon began to be understood.

Henry's work contained the germ of yet another important discovery. Some of his experiments on induction by Leyden jar discharges involved the transmission of electric force without wires through distances as great as two hundred feet, and through the floors and walls of buildings. And in similar experiments in which he

observed the effects of lightning flashes in place of sparks from a Leyden jar, he found that he could get the lightning to magnetize needles up to a distance as great as eight miles. This was about 1842. Here we have the earliest evidence of ether waves of the type that the radio engineer employs. But again the significance of Henry's work was not recognized. This could only have come after much fuller investigation. However, it is instructive to reflect for a moment on what might have been had Henry possessed the time and facilities for carrying his work further. Needless to say, there is a wide gulf between the wireless telegraph of today and its earliest precursor with which Henry received an electromagnetic signal from a lightning flash eight miles away, but it is wholly possible that, had Henry not been called to other work, the world might have possessed a wireless telegraph capable of sending messages over substantial distances many years before it did.

Writing of Henry, Simon Newcomb, the celebrated astronomer said,[2] "His scientific work is marked by acuteness in cross-examining nature, a clear appreciation of the logic of science, and an enthusiasm for truth without respect to its utilitarian results." A man of the highest scientific ability, Henry spent the better part of his life as the head of an institution dedicated to "the increase and diffusion of knowledge among men."

"The mantle of Franklin has fallen upon the shoulders of Henry," wrote Sir David Brewster,[3] the eminent English scientist, and it is reported that Abraham Lincoln declared, when he became acquainted with Henry after assuming the Presidency, "The Smithsonian Institution must be a grand school if it produces such thinkers as Henry." He was, in every way and in the best that the word implies, a scientist, and the interest in scientific questions which dominated his life, remained with him to the very end,—almost the last words to pass his lips were whether the transit of the planet mercury had been successfully observed. If we use the word "Dean"—so rich in academic association—to stand at once for the greatest usefulness to one's fellowmen as well as for the highest achievements in the field of scholarship and research, for lifelong devotion to public service, for breadth of view and tolerance regarding all questions, whether arising in science or directly out of human relations, and as epitomizing all that is best and highest in man's intellectual life, we may well call Joseph Henry the Dean of American scientists.

[2] Biographical Memoir; National Academy of Sciences, Apr. 21, 1880.
[3] Biographical Memoir; prepared by Prof. Asa Gray in behalf of the Board of Regents of the Smithsonian Institution.

Correction of Data for Errors of Measurement

By W. A. SHEWHART

Introduction

EVERY measurement is subject to error. This universally accepted truth is the result of every-day experience. From the simplest type of measurement, such as determining the length of a board with an ordinary tape measure, to the most refined type of measurement, such as determining the charge on an electron, errors are bound to creep in.

Now, a manufacturer must constantly make measurements of one kind or another in an effort to control his production processes and to measure the quality of his finished product in terms of certain of its characteristics, but, before he can safely determine the significance of observed differences in his production processes or in the quality of his product as given by these measurements, he must make allowance for his errors of measurement; i.e., for the fact that the observed differences may be larger or smaller than the true differences. To make such allowances for the errors of measurement of any characteristic, to find out what the true magnitude of the characteristic most probably is, to find out, as it were, what a thing most probably is from what it appears to be, presents an endless chain of interesting problems to be solved.

Three important types of problems arising in engineering practice are discussed in this paper. They are:

1. Error correction of data taken to show the quality of a particular lot.

2. Error correction of data taken periodically to detect significant changes in quality of product.

3. Error correction of data taken to relate observed deviations in quality of product to some particular cause.

The solution of the first one is presented here for the first time. The solution of the second has been generalized to include cases not previously solvable. All three types of problems are illustrated.

PART I

Type 1—Error Correction of Data Taken to Show the Quality of a Particular Lot

Let us take a specific problem first. Assume that we have a lot consisting of 15,000 transmitters [1] and a machine with which to measure the efficiency of each instrument. Suppose we make one observation on each transmitter—a total of 15,000 measurements. Suppose we find, as in the distribution illustrated in Fig. 1, that one measurement is in the efficiency range -1.75 to -1.50, 17 within the range

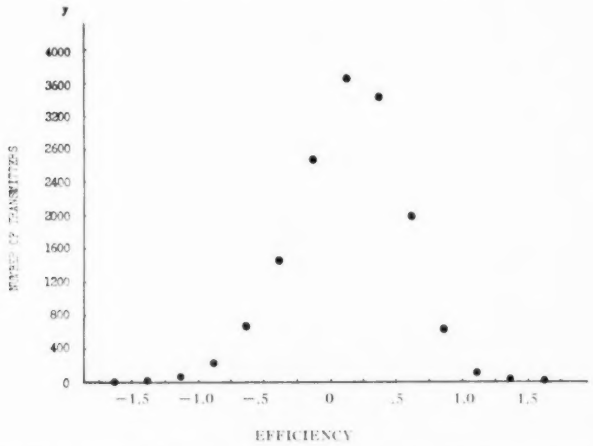


Fig. 1—Typical frequency distribution. Chart showing observed number of transmitters versus efficiency

-1.50 to -1.25 units, and so on. The vertical height of a point represents the number or frequency of occurrence of observations falling within the corresponding interval laid off on the horizontal axis of the chart.

So far so good, but suppose a customer wants to buy these transmitters. We know that some transmitter which appeared to have an efficiency within the range of 1.25 to 1.50 units say, *may* actually have had an efficiency within some other interval. We know too that, because of the errors of measurement, the transmitters appear to differ more among themselves than they really do. We therefore

[1] Of course, the efficiency of a transmitter does not remain constant during a series of tests but these inherent variations in the transmitter may be considered, for our purpose, as forming a component part of the resultant error of measurement.

desire to find the most probable numbers of transmitters within the different intervals indicated in Fig. 1.

Analytical Statement of Problem

Let us assume that the most probable number of transmitters within the interval of efficiency from X to $X+dX$ is $f_I(X)dX$. It is this function $f_I(X)$ that we want to find. Similarly let us assume that there is some function $f_O(X)$ such that $f_O(X)dX$ gives the observed number of transmitters appearing to have efficiencies within the interval X to $X+dX$ where the measurements are made by a method wherein the probability of making an error within the interval x to $x+dx$ is $f_E(x)dx$. It is reasonable to expect that, if two of these functions are known, the third can be easily determined. We shall proceed to show that this is the case. Let us first find the law of error experimentally.

Finding the Law of Error

The problem is to determine the chance of making an error of a given magnitude in measuring the efficiency of any transmitter. Naturally, the only way of doing this is to make a series of measurements on a single transmitter from which we can determine the observed frequency of occurrence of measurements which differ from the average by some fixed amount, and thus find what percentage of the total number of measurements may be expected to fall within any given range on either side of the average. Common sense and intuition may tell us that we may expect to find a large percentage of the measurements within a narrow range on either side of the average, that there will be just as many measurements greater than the average by a certain amount as there are less than the average by the same amount, and that large deviations from the average may be expected to occur with less frequency than small deviations. Suppose we make 500 observations of the efficiency of a single transmitter and find the distribution given in Fig. 2. Just as we might have expected, the observed values of the efficiency of the transmitter are grouped symmetrically about the average of all the observed values. We see that the maximum deviation between observations on a single transmitter is quite large (33%) compared with the actual maximum differences observed between the efficiencies of the transmitters.

The results reproduced in Fig. 2 suggest that the deviations for the case in hand are distributed in a manner closely approximating the

bell-shaped distribution so familiar in the theory of errors. We often find, as we do in this case, that the observed distribution can be closely approximated by a function $f_E(x)$ of the form

$$f_E(x)dx = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-\bar{X})^2}{2\sigma^2}} dx, \tag{1}$$

where $f_E(x)dx$ is the probability that an error x will lie within the interval x to $x+dx$, σ is the root mean square or standard deviation, $\bar{X}$ is the arithmetic mean value and $(X-\bar{X})$ is the deviation x. The

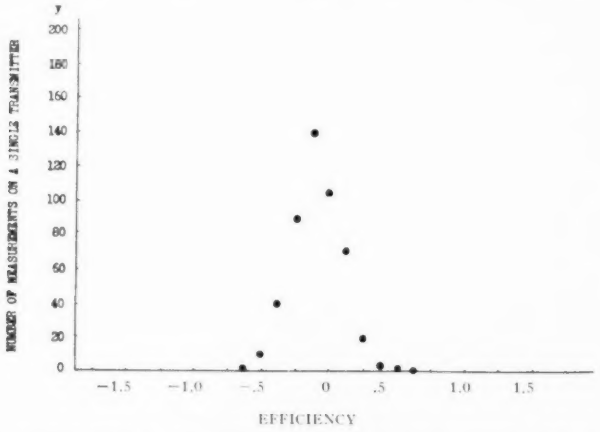


Fig. 2—Typical form of distribution of errors of measurement. Chart showing number of measurements on a single transmitter versus efficiency

function $f_E(x)$ is referred to in the literature as the normal law of error. If we try to fit such a curve to the deviations[2] given in Fig. 2, we obtain the results shown in Fig. 3. This figure is the same as Fig. 2 except for the addition of the smooth normal curve of error calculated for the observed data. Without further consideration, we shall assume the law of error to be normal and hence of the form indicated by Equation (1).

Finding the True Distribution $f_T(X)$

We have next to consider the choice of the function to represent the true distribution $f_T(X)$. Often we have reason to believe that this

[2] If the average of the observed values of the 500 observations of efficiency given in Fig. 3 is assumed to be the true value of the efficiency of the transmitter, then the deviation of an observed value from this mean is also the error of this observed value. We shall use the terms "error" and "deviation" interchangeably in this sense.

is also approximately normal, and hence we shall consider first the method for finding the observed distribution $f_o(X)$ for the special case when both the true distribution $f_T(X)$ and the law of error $f_E(X)$ are normal; i.e., when they are both of the form given by Equation (1).

We shall first obtain an experimental answer to this problem. Suppose we take, say, 1,000 instruments of some kind which are

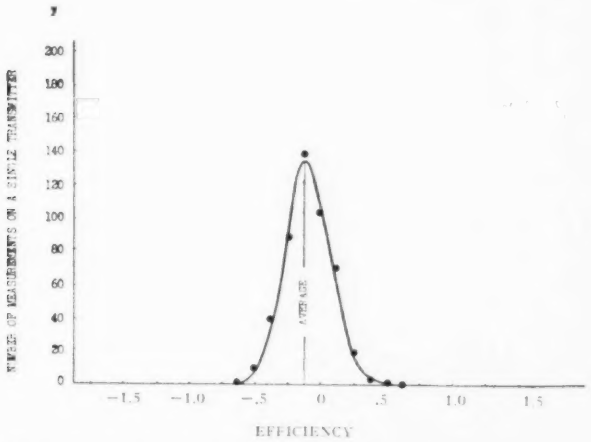


Fig. 3.—Chart showing the observed distribution of errors fitted by a typical smooth curve. Data of Fig. 2 fitted by normal law of error, Eq. 1

known to be distributed in normal fashion, in respect to some characteristic, with a standard deviation σ_T. Let us measure each of these instruments by a method subject to the normal law of error whose standard deviation σ_E is $\frac{1}{2}\ \sigma_T$. The results of one such experiment are given in Fig. 4. The observed frequencies of occurrence are represented by the circles. It was found that this observed distribution could be closely approximated by a normal law $f_o(X)$ for which the standard deviation σ_o was $\sqrt{\sigma_T^2+\sigma_E^2}$. This experiment suggests a general theorem which will be demonstrated analytically in a succeeding paragraph. The theorem is: When the true distribution $f_T(X)$ and the law of error $f_E(x)$ are both normal (hence expressible in form indicated by Equation (1)) with root mean square or standard deviations σ_T and σ_E respectively, the most probable observed distribution will be normal in form with a standard deviation $\sigma_o=\sqrt{\sigma_T^2+\sigma_E^2}$.

The observed distribution in Fig. 1 is asymmetrical and hence not

normal as it should be if $f_T(X)$ and $f_E(x)$ were both normal. We must therefore, try some other function for $f_T(X)$.

Of course, experiments might be performed for other types of true and error distributions, but in all such cases the results, as in the illustration just considered, would be subject to errors of sampling.

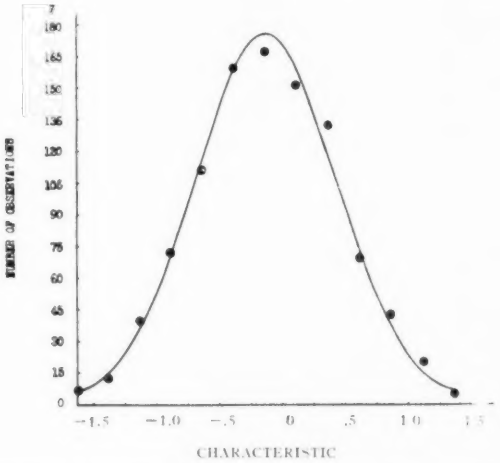


Fig. 4—Experimental results shpwing effects of errors of measurement. Normal curve fitted to observed points, when the true distribution and the law of error are both normal

Hence we shall proceed at once to the analytical treatment of the problem.

Assuming the law of error to be normal, we see that the fraction $f_E(x)dx$ of the number of objects having magnitudes between $X+x$ and $X+x+dx$ will be measured with an error between $-x$ and $-x-dx$ and hence will be observed as of magnitude X (Fig. 5). Thus

$$f_o(X) = \int_{-\infty}^{\infty} f_T(X+x) \frac{1}{\sigma_E \sqrt{2\pi}} e^{-\frac{x^2}{2\sigma_E^2}} dx. \tag{2}$$

For the particular case treated in a previous paragraph where both the true distribution $f_T(X)$ and the law of error $f_E(x)$ are normal, we may write Equation (2) in the form

$$f_o(X)dX = \frac{1}{\sigma_T \sigma_E 2\pi} \int_{-\infty}^{\infty} e^{-\frac{(X+x)^2}{2\sigma_T^2}} e^{-\frac{x^2}{2\sigma_E^2}} dX\, dx \tag{3}$$

where σ_T and σ_E are the root mean square or standard deviations of

the true and error distributions respectively. Integration of Equation (3) gives [3]

$$f_o(X) = \frac{1}{\sigma_o \sqrt{2\pi}} e^{-\frac{X^2}{2\sigma_o^2}}, \tag{4}$$

where

$$\sigma_o = \sqrt{\sigma_T^2 + \sigma_E^2}. \tag{5}$$

Equations (4) and (5) are the analytical expression for the rule stated previously, for finding the observed distribution $f_o(X)$ when both the true and error distributions are normal, because Equation (4)

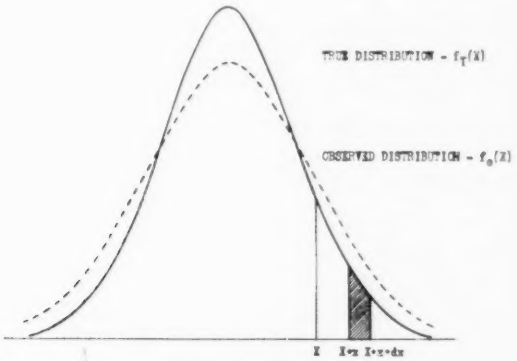


Fig. 5—Chart used in explaining the derivation of $f_o(X)$ in terms of $f_T(X)$

shows it to be normal and Equation (5) expresses the standard deviation σ_o of the observed values in terms of those of the true values and of the errors.

In practice, however, we often find that the true distribution is non-symmetrical or skew and can be more nearly approximated by the function [4]

$$f_T(X) = \frac{1}{\sigma_T \sqrt{2\pi}} e^{-\frac{X^2}{2\sigma_T^2}} \left[1 - \frac{k_T}{2} \left(\frac{X}{\sigma_T} - \frac{X^3}{3\sigma_T^3} \right) \right] \tag{6}$$

where k_T is a measure of the asymmetry or skewness, the modal or most probable value of X being at a distance $-\frac{k\sigma_T}{2}$ from the average

[3] See Appendix 1 where another method of solution is given.

[4] This is often referred to in the literature of statistics as the second approximation. It is in fact the first two terms of the Gram-Charlier series.

value of X. Substitution of this expression and a normal error function in Equation (2), yields upon integration [5] the following distribution $f_o(X)$ of the observed values

$$f_o(X) = \frac{1}{\sigma_o \sqrt{2\pi}} e^{-\frac{X}{2\sigma_o^2}} \left[1 - \frac{k_o}{2} \left(\frac{X}{\sigma_o} - \frac{X^3}{3\sigma_o^3} \right) \right] \tag{7}$$

where

$$\sigma_o = \sqrt{\sigma_T^2 + \sigma_E^2}, \tag{5}$$

and

$$k_o = k_T \frac{\sigma_T^3}{\sigma_o^3}. \tag{8}$$

We see that the distribution $f_o(X)$, Equation (7), of the observed values is of the same form as that $f_T(X)$, Equation (6), of the true values. The standard deviation of the errors of measurement σ_E, as in the previous case, has equal weight with the standard deviation σ_T in influencing the standard deviation σ_o of the observed values. The degree of asymmetry of the observed distribution as measured by the skewness k_o is, however, less (Equation (8)) than that of the true distribution as measured by the skewness k_T of the true distribution.

Now we can correct the observed distribution, Fig. 1, for the errors of measurement, because we find that the observed frequencies,Fig. 1, can be closely approximated by a function of the type defined by Equation (7). Knowing that the law of error, Fig. 3, is normal we conclude that the true distribution $f_T(X)$ must be a function of the same type as $f_o(X)$ was found to be except that the true standard deviation σ_T will be, from Equation (5), $\sqrt{\sigma_o^2 + \sigma_E^2}$ and the true skewness k_T will be, from Equation (8), $\frac{\sigma_o^3}{\sigma_T^3} k_o$. Now, σ_o and k_o can be calculated from the observed distribution, Fig 1, and σ_E can be determined by the data given in Fig. 3.

Thus finding the values of σ_T and k_T and substituting them in Equation (6), we have the function $f_T(X)$ representing the true distribution which we started out to find. From this knowledge of $f_T(X)$ we can now get the most probable frequencies of occurrence of the different efficiencies. Subtracting these frequencies from those observed and shown in Fig. 1, we get the corrections plotted in Fig. 6, expressed as percentages of the observed frequencies.

[5] This solution is also obtained by another method in Appendix 1.

Summary

We are now in a position to summarize the practical routine to be followed in finding the most probable distribution $f_T(X)$ of quality when the observed distribution is given.

To find $f_T(X)$, we must first know the law of error $f_E(x)$. We must show this to be normal and find the standard deviation σ_E

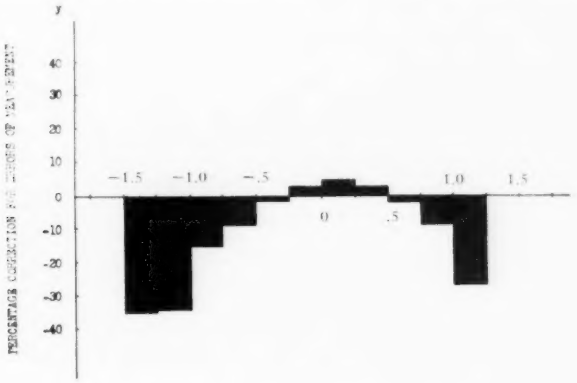


Fig. 6—Correction which must be applied to the observed distribution of transmitters Fig. 1, because of the existence of errors of measurement

by multiple tests on a single unit. The error made in determining the standard deviation σ_E from n observations is $\frac{\sigma_E}{\sqrt{2n}}$. Hence the precision we attain in finding $f_T(X)$ depends upon the number of observations n made in finding σ_E.

Having found σ_E to the required degree of precision, we must next discover whether or not the observed distribution $f_T(X)$ is either normal or the second approximation. Standard statistical methods can be used for this purpose.

If the $f_o(X)$ is normal, we then know that

$$f_T(X) = \frac{1}{\sqrt{2\pi(\sigma_o^2 - \sigma_E^2)}}\, e^{-\frac{X^2}{2(\sigma_o^2 - \sigma_E^2)}},$$

and, if $f(X)$ is second approximation, we know that $f_T(X)$ is given by Equation (6), where σ_T and k_T can be found with the aid of Equa-

tions (5) and (8) in terms of the observed values of σ_E, σ_o and k_o. In other words we have

$$f_T(X) = \frac{1}{\sqrt{2\pi(\sigma_o^2 - \sigma_E^2)}} e^{-\frac{X}{2(\sigma_o^2 - \sigma_E^2)}} \left[1 - \frac{k_o \sigma_o^3}{(\sigma_o^2 - \sigma_E^2)^{\frac{3}{2}}} \left(\frac{X}{(\sigma_o^2 - \sigma_E^2)^{\frac{1}{2}}} - \frac{X^3}{3(\sigma_o^2 - \sigma_E^2)^{\frac{3}{2}}} \right) \right].$$

PART II

Correction of Data Taken Periodically to Detect Significant Changes in Quality of Product

Irrespective of the care taken in defining and controlling the manufacturing processes, the units of a product will differ among themselves in respect to any measurable characteristic. Random fluctuations in such factors as humidity, temperature, grade of raw material, and wear and tear on machinery may produce such differences between units of a product. Such random variations in the factors underlying the manufacturing process usually yield a product in which the units differ in random fashion according to some law of probability.

Customarily, product is inspected periodically, and the data are analyzed to determine if the observed difference in two samples is greater than can be accounted for as a random variation. If it is, we may assume that the manufacturing processes have changed significantly for some reason which further investigation should disclose. Now, the presence of errors of measurement effectively increases the magnitude of the random differences to be expected from one sample to another and hence makes it harder for us to detect trends or fluctuations in product. Let us investigate this effect of errors of measurement.

Symbolic Statement of Problem

Symbolically we may assume that the probability of production of a unit of product having a characteristic X within any range X to $X + dX$ is $f_T(X)dX$, where the characteristic X is measured by a method subject to a law of error $f_E(x)$, so that $f_E(x)dx$ represents the probability of occurrence of an error x within the range x to $x+dx$. The problem is to find the corresponding distribution $f_c(X)$ for the observed magnitudes.

General Solution of Problem

Obviously the observed magnitude X_o is the algebraic sum of the true value X and the error x. Assuming that there is no correlation between these two quantities, the probability of a unit having a value of X within the range X to $X + dX$ being measured with an error x within the range x to $x + dx$ is $f_I(X)dX\ f_E(x)dx$. Assuming that $X_o = X + x$ we may write the probability

$$y_o = f_o(X_o)dX_o = \int_{-\infty}^{\infty} f_I(X_o - x)dX_o f_E(x)dx,$$

because $f_o(X_o)$ is obtained by taking into account that all possible values of x between $+\infty$ and $-\infty$ may be combined with a given X. This integral is of the same form as that given in Equation (2). Integration for the case where both $f_I(X)$ and $f_E(x)$ are normal gives

$$f_o(X_o) = \frac{1}{\sigma_o \sqrt{2\pi}} e^{-\frac{X_o^2}{2\sigma_o^2}}$$

where as before $\sigma_o = \sqrt{\sigma_I^2 + \sigma_E^2}$. This result is well known as the law of propagation of error.

When $f_E(x)$ is normal and $f_I(X)$ is given by the first two terms of the Gram-Charlier series, Equation (6), with skewness k_I and standand deviation σ_I, the observed distribution $f_o(X_o)$ is of the same functional form as the true distribution $f_I(X)$ and has values of standard deviation σ_o and skewness k_o given by Equations (5) and (8) in Part I. This result appears to be new.

Now for the case where the true distribution $f_I(X)$ and the law of error $f_E(x)$ are both second approximation type, the integration is somewhat tedious, but we can approach a special case of this problem easily from a slightly different angle as indicated in Appendix 2. Under certain special conditions therein set forth, the resultant distribution is also second approximation form with a skewness which is less than that of either $f_I(X)$ or $f_E(x)$ and is equal to $\frac{1}{\sqrt{2}} k_I$ when $k_I = k_E$, the standard deviation σ_o being again equal to $\sqrt{\sigma_I^2 + \sigma_E^2}$.

Example of Applications to Determine Most Economical Way of Measuring Quality

Let us next consider a very simple method of using the above results to indicate the most economical method for determining the quality of product with a given degree of precision.

What is the most economical way of determining the quality of product within some predetermined range $\overline{X} \pm \Delta\overline{X}$ with a known probability P, where $\overline{X}$ is the average quality? Let us assume that:

a_1 = cost of selecting each unit and making it available for measurement,

a_2 = cost of making each measurement,

n_1 = number of units selected,

n_2 = number of measurements made on each unit,

σ_1 = standard deviation of the errors of observation,

$\sigma_2 = \sigma_I$ = standard deviation of the true distribution $f_I(X)$.

Let us take $P = .9973$. Then the range $X \pm 3\sigma_X$ includes 99.73 per cent. of the observations, and hence $\Delta X = 3\sigma_X$.

The average of n_2 measurements made on one unit is the observed value of the magnitude X for that unit, and this average has the standard deviation $\sigma_E = \frac{\sigma_1}{\sqrt{n_2}}$. Hence, from the theory of the preceding section, the standard deviation of the observation is

$$\sigma_o = \sqrt{\sigma_I^2 + \sigma_E^2} = \sqrt{\sigma_2^2 + \frac{\sigma_1^2}{n_2}}.$$

The standard deviation of the average of n_1 observations is $\sigma_{\overline{X}} = \frac{\sigma_o}{\sqrt{n_1}}$ and we find upon solving for n_1,

$$n_1 = \frac{\sigma_2^2 + \frac{\sigma_1^2}{n_2}}{\sigma_{\overline{X}}^2}.$$

The cost of inspection is

$$y = a_1 n_1 + a_2 n_1 n_2,$$

and by customary methods this can be shown to be a minimum when

$$n_2 = \frac{\sigma_1}{\sigma_2}\sqrt{\frac{a_1}{a_2}}.$$

The following values correspond to one practical case:

$\Delta\overline{X} = .3$ unit	$a_1 = \$0.50$
$\sigma_1 = .3$ unit	$a_2 = \$0.02$
$\sigma_2 = .9$ unit	$P = .9973$

Thus with the aid of the above theory we find the most economical method of inspection requires 2 observations on each of 86 units.

Application in Setting Limit Lines

Over 99 per cent. of the averages of samples of size N drawn from a product whose law of distribution is $f_I(X)$ where $f_I(X)$ is either normal or second approximation may be expected to lie within the limits defined by the true average $\overline{X}$ plus or minus $3\frac{\sigma_T}{\sqrt{N}}$. If an average falls outside these limits, this fact is taken as probably indicating the existence of a trend or cyclic fluctuation in product, the cause of which should be sought. The presence of errors of measurement increases the separation of these limits to $6\sigma_m$ from $6\sigma_T$. Our precision of detecting trend or cyclic fluctuation is thereby decreased.

Cases often happen in practice where σ_m is from 15 per cent. to 25 per cent. greater than σ_T. In some instances σ_m has been found to be nearly 50 per cent. greater than σ_T.

PART III

Error Correction of Data Taken to Relate Observed Deviations in Quality of Product to Some Particular Cause

In many practical cases it is not possible to write down an equation to show how the quality of a finished product depends upon the factors controlled by different manufacturing steps. To cite one such case, we may know that the quality of the finished article depends upon the control of the temperature to which some of the piece parts are heated in the process of manufacture. Thus the microphonic properties of carbon depend upon the temperature to which the carbon is heated. In cases where the relationship between quality and some factor (such as temperature in the above illustration) can only be determined through a study of the correlation existing between the quality and the particular factor, use must be made of the correlation coefficient r which is defined as

$$r = \frac{\Sigma\, y\, x}{\sigma_x \sigma_y N}$$

where x and y represent respectively deviations from the average quality $\overline{X}$ and the average magnitude Y of some factor which is

to be controlled by the manufacturing process, and N is the number of observations. Now, if errors of observation are made in determining x and y, the observed correlation coefficient $r_{x_0y_0}$ is known to be given by the expression

$$r_{x_0y_0} = \frac{\sigma_x \sigma_y}{\sigma_{x_0} \sigma_{y_0}} r_{xy} \tag{10}$$

where $\sigma_{x_0} = \sqrt{\sigma_x^2 + \sigma_{x_E}^2}$ and $\sigma_{y_0} = \sqrt{\sigma_y^2 + \sigma_{y_E}^2}$,

σ_{x_E} and σ_{y_E} being the root mean square errors of observation of x and y respectively.

Attention is directed to Equation (10) which shows that the observed correlation coefficient $r_{x_0y_0}$ is always less than the true correlation coefficient r_{xy} irrespective of the number of observations made. Obviously, this point is of considerable commercial importance as we shall now see.

If the observed correlation is small, we customarily assume that there is little need of trying to control the quality X by controlling the manufacturing factor Y, whereas this conclusion cannot be justified unless it can be shown that the true correlation has not been masked by the errors of measurement.

This point has had to be taken into account in the development of machine methods for testing transmitters and receivers, because the calibration curves of the machines in terms of ear-voice tests depend upon the correlation coefficient.

Appendix I

It may be of some interest to certain readers to note that the results given in Equations (4) and (7) can also be obtained in the following way by the method of moments so often used in statistical investigations.

Assuming that $f_T(X+x)$ is expansible in terms of a Taylor's series, we get

$$f_o(X) = f_T(X) + \frac{\sigma_E^2}{2} f_T''(X) + \frac{1}{\lfloor 2} \left(\frac{\sigma_E^2}{2}\right)^2 f_T''''(X) + \frac{1}{\lfloor 3} \left(\frac{\sigma_E^2}{2}\right)^3 f_T^{vi}(X) + \ldots . \tag{11}$$

If we substitute a normal form for $f_T(X)$ in Equation (11) and solve for the moments of $f_o(X)$, we find that the odd moments are zero

and the ratio of the 4th moment to the square of the 2nd is numerically 3 which indicates that $f_o(X)$ is normal in form.

A similar substitution of the 2nd approximation form for $f_T(X)$ in Equation (11) yields a distribution $f_o(X)$ from whose moments we deduce Equation (7). Use is made in this proof of the easily demonstrated theorem that

$$\int_{-\infty}^{\infty} x^i f_E^j(x) = 0$$

if $i < j$, where f_E^j is the jth derivative of the normal law function.

APPENDIX II

It is well known that the normal law of distribution may result from a system of n (n being large) causes each of which produces an increment ΔX measured from some fixed origin with a probability $p = \frac{1}{2}$ and no increment with a probability $q = \frac{1}{2}$. Furthermore the second approximation may result from a similar system in which $p+q$ and n is large. Under such systems of causes, the probabilities of the occurrences of n, $n-1$, . . . 3, 2, 1, 0 increments are given by the successive terms of the point binomial $(p+q)^n$.

Let us assume that the symbols p_T, q_T, n_T, ΔX and p_E, q_E, n_E, Δx refer to the systems of causes controlling the product and errors respectively. The probabilities of observed combinations $n_T\Delta X + n_E\Delta x$, $(n_T-1)\Delta X + (n_E-1)\Delta x$, . . . are given by the successive terms of the expansion $(p_T+q_T)^{n_T}\ (p_E+q_E)^{n_E}$. Now for the special case $p_T = p_E = p$ and $\Delta X = \Delta x$ we have the resultant probability distribution $(p+q)^{n_T+n_E}$ with skewness

$$k_o = \frac{q-p}{\sqrt{pq(n_T+n_E)}}$$

and standard deviation

$$\sigma_o = \sqrt{pq(n_T+n_E)}.$$

Now if $p = q$, the skewness k_o is zero and the observed distribution is more nearly normal than either component, and its standard deviation σ is the square root of the sum of the squares of σ_T and σ_E. This result is similar to that given by Equation (4) of this paper.

We may also consider by this method a case not treated in this paper. When the skewness k_T of the true values is equal to that k_E of the law of error, or, more particularly, when $n_T = n_E = n$, $p_T = p_E = p$, $q_T = q_E = q$, $p = q$, we see that the observed distribution is given

by the successive terms of $(p+q)^n$ and the skewness of the observed distribution k_o is $\frac{1}{\sqrt{2}}\,k$, and the standard deviation σ_o is $\sqrt{2}\,\sigma$; i.e. the observed skewness is only $\frac{1}{\sqrt{2}}$ times that of either the true distribution or the law of error, and the observed standard deviation σ_o is $\sqrt{2}$ times the standard deviation of either of the true or error distributions.

The Theory of the Operation of the Howling Telephone with Experimental Confirmation

By HARVEY FLETCHER

Synopsis: A general theory of the sustained oscillations of electro-mechanical systems is presented in the paper. The electrodynamical properties of the telephone transmitter and receiver are described and sufficient numerical data are given to enable one to calculate the intensity and frequency of howling for various types of systems. Detailed consideration is given to the following three systems, namely, one where the transmitter and receiver disphragms are coupled together mechanically by a lever system, one where they are coupled by a small box of air, and one where they are coupled by a long tube of air. The type of electrical circuit to use with each of these systems depends upon the type of performance desired.

WHEN the telephone receiver of a subscriber's set is held in front of the mouthpiece of the transmitter, a shrill note is emitted. A sustained oscillation is set up in the electro-mechanical system which is frequently called "howling" or "singing" or "humming."

This phenomenon was first observed by A. S. Hibbard of the United States in 1890. Frank Gill was the first to publish an account of the phenomenon. He first noted that the pitch of the howling note was changed by reversing the telephone receiver connection. In summarizing further his experimental results, he states "that the pitch of the note appears to be determined by the length of the column of air between the two diaphragms and the conditions of the circuit. As the periodic time of the circuit is increased, the time of the note rises. To some extent, the pitch is governed by the rate of the diaphragm, but I do not think this is so important a factor as the others. The main factors appear to be the angle of lag and the length of the column of air between the diaphragms. Although the vibration is a forced one, we could almost see that its rate is largely dependent on the free period of the circuit." [1]

In 1908 Kennelly and Upson extended Gill's work and made extensive experimental investigations of the case in which the transmitter and receiver are coupled together acoustically by means of a

[1] Taken from a paper on "Notes on the Humming Telephone" by F. Gill, read at a meeting of the Dublin Local Section of the Society of Telephone Engineers and published in the Journal of the Institution of Electrical Engineers, Vol. XXXI, 1901.

hollow circular tube of varying lengths and electrically by means of an induction coil. The summary of the conclusions is as follows: [2]

"(1) The mean frequency of the humming-telephone note is determined solely by the receiver diaphragm, and its natural free rate of vibration. (2) The ascending intersections of the frequency zig-zag with the mean frequency line will be formed approximately at tube lengths of $(3/4+m)\ v/n_o$ cm. for one connection, and of $(1/4+m)\ v/n_o$ cm. for the other connection, of the receiver; where v is the velocity of sound in air, n is the mean frequency in cycles per second, and m is any positive integer, within the working range of the tube. The constants 3/4 and 1/4 may be modified by the presence of condensers, and other circumstances. (3) The range of pitch variation, and the breaking positions, are determined by the transmitter, and by the reinforcing capability of the system. For systems that are weak, either electrically or acoustically, the range of pitch, above or below the mean, will be small. (4) The primary current, as measured by a DC instrument, is ordinarily a minimum at the mean frequency, and a maximum at a break. (5) Transmitters may be tested for effectiveness, by measuring their hum-extinguishing resistances in the primary or secondary circuit. The tube length should be such as to produce mean frequency if one connection of receiver only is used, but should favor both connections equally, if both connections of receiver are used."

They also give a first approximation theory to account for the changes in frequency as the length of the coupling tube is changed.

In 1917, H. W. Nichols gave the general equations for the special case where the two diaphragms act as pistons closing the ends of a tube of air. This case was given as an illustrative example of the "Theory of Variable Dynamical Electrical Systems." [3]

This paper gives a theoretical treatment of the behavior of a system containing a transmitter and a receiver coupled together acoustically and electrically, and with a source of electrical energy feeding the transmitter. Formulae are deduced which give the frequency and intensity of howling in terms of the physical constants of the system. Numerical calculations are given and sufficiently detailed solution of some special cases are given to enable one, who is interested in using the howling telephone as a source of alternating current or for other experimental work, to design the set for his particular purpose.

[2] "Humming Telephone" by A. E. Kennelly and Walter L. Upson, American Philosophical Society, July 20, 1908.

[3] Physical Review, Aug., 1917, p. 191.

General Solution of the Howling Circuit

The elements of a telephone system which is howling are the transmitter, the receiver, the mechanical coupler and the electrical coupler as indicated in Fig. 1. If there is a source of electrical power in the electrical coupler, which is released by movements of the transmitter diaphragm in the form of electrical vibrations, and also, if there is a proper relationship between these four elements, then a sustained

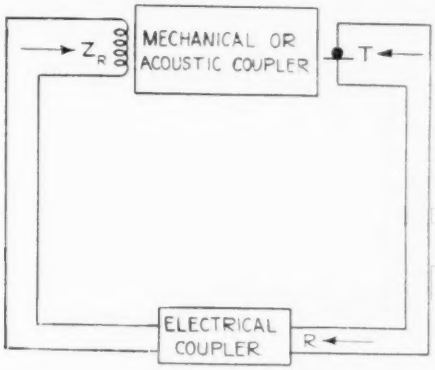


Fig. 1

howling will result. In other words, if the gain in the transmitter due to its amplifying action is just equal to the losses in the electrical and mechanical circuits, then a steady oscillatory state will be maintained. The problem is to determine the nature of these relationships.

Assume that the conditions are such that a steady oscillatory state has been set up. Under such conditions let T be the electrical impedance of the transmitter, R the impedance looking away from the transmitter terminals into the electrical coupler, and Z_R the impedance of the receiver. It is well known that the impedance Z_R is dependent upon the velocity of motion of the receiver diaphragm. Also, T is dependent upon the amplitude of motion of the transmitter diaphragm as well as upon the direct current supplied to it. Consequently, the impedances defined above are not only dependent upon frequency but also upon the mechanical coupling and magnitude of the current supplied to the transmitter.

If e is the electromotive force created in the transmitter, and i the

current flowing through it both expressed in root mean square values[4], then

$$e=(T+R)i \tag{1}$$

It is convenient to define a quantity M which I shall call the unilateral mutual impedance by the equation

$$e_1=Mi_1 \tag{2}$$

where e_1 is the electromotive force created in the transmitter when a current i_1 flows in the receiver circuit. It is a quantity which is closely related to the effectiveness of the mechanical coupling and the efficiencies of the transmitter and receiver.

If the electrical coupler be considered part of the receiver, and the transmitter and receiver circuits are connected together as in Fig. 1, then $e=e_1$, and $i=i_1$. Consequently

$$M=T+R \tag{3}$$

is the condition for sustained oscillation. This condition is in effect a pair of conditions, as the two sides of the equation must be equal both in amplitude and in phase. These two conditions are sufficient to determine the frequency and intensity of howling.

In order to express M and R in more fundamental physical constants, it is necessary to examine more closely the mechanical and electrical connections. Before doing this for some important special cases, it will be necessary to discuss some of the electro-dynamical properties of transmitters and receivers.

Electrodynamical Properties of Transmitters and Receivers

For the sake of clarity the discussion will be confined to permanent magnet receivers and carbon transmitters. The modifications necessary for other types of instruments will, I think, be evident from the discussion. Representing by F_R and F_T the forces acting on the diaphragms of the receiver and transmitter respectively, and by y and z their displacements, we have the following equations defining the "stiffness factors" S_R and S_T

$$S_R=\frac{F_R}{y} \tag{4}$$

$$S_T=\frac{F_T}{z} \tag{5}$$

[4] In what follows all quantities involving periodic variations will be expressed as root mean square values unless otherwise specified, and the vector notation will be used for denoting phases.

These factors are usually complicated functions of the frequency while S_T likewise depends on the kind and amount of agitation. In the case of a system of a single degree of freedom which may be regarded as a first approximation to this case

$$S = m\omega^2 + j\omega r + s \tag{6}$$

where ω is 2π times the frequency. When referring to the movements of a diaphragm, the quantity m represents the mass, r the mechanical resistance, and s the elastic constant. The stiffness factor S divided by $j\omega$ is usually called the mechanical impedance.

Measurements have shown that for the transmitters and the

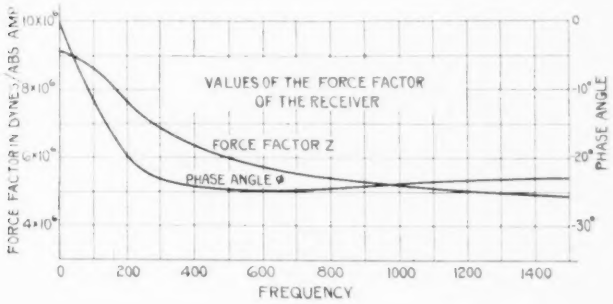


Fig. 2

receivers used in the experiments described below, the following constants represent approximately the two stiffness factors in the region of resonance

$$S_R = -.93\omega^2 + 230j\omega + 3\times 10^7 \tag{6'}$$

$$S_T = -4.5\omega^2 + 2000j\omega + 2\times 10^8 \tag{6''}$$

An important constant which enters into the determination of the unilateral mutual impedance M is the force factor of the receiver which will be designated by Z. It is defined as the force in dynes acting upon the diaphragm per unit of current. For the receivers used in this investigation, its values in magnitude and phase are shown for various frequencies in Fig. 2. These were determined by the method outlined by Wegel.[5] In the region of the resonant frequency its value in absolute units can be approximately represented by

$$Z = 5.3\times 10^6 \,\angle\, 24°. \tag{7}$$

[5] Theory of Telephone Receivers—Wegel, R. L., Jour. of A. I. E. E., Oct. 1921.

The impedance Z_R of the receiver varies with frequency and depends upon the load on the diaphragm. If S is the loaded stiffness of the diaphragm, that is, its resistance to force under actual working conditions, and Z_d is the impedance of the receiver when the diaphragm is prevented from moving, then it is well-known that

$$Z_R = Z_d + j\frac{\omega Z^2}{S}. \tag{8}$$

It was found that Z_d expressed in ohms could be represented in the frequency region near resonance by the formula

$$Z_d = 93 + .06f + j(43 + .15f) \tag{9}$$

where f denotes the frequency in cycles per second.

The electromotive force e created in the transmitter, the direct current I flowing through it, and the displacement of the diaphragm are related in a rather complicated way. For describing this relationship it is convenient to define a modulation factor h by the equation

$$e = Ihz \tag{10}$$

Combining this equation with (2) it is seen that

$$M = Ih\frac{z}{i} \tag{11}$$

which shows that the modulation factor is also an important one in determining the unilateral mutual impedance. For a sustained oscillation the factor Ih does not enter into the periodic variation and may be thought of as an electro-mechanical impedance between the electromotive force created in the button and the displacement of the diaphragm of the transmitter. However, for a different condition of sustained oscillation which results in giving z a different magnitude the value of h changes. In other words h is dependent upon the agitation of the carbon as represented by z, and also upon the direct current supplied to the transmitter. It is mainly this variable character of h that makes it possible to fulfill the conditions for sustained howling.

Simultaneous measurements of e, I and z were made upon several transmitters of the type used in this investigation. From the results obtained and from the defining equation (10) for h, it was found that

the following empirical equation would represent approximately the relation between h, I and z, namely

$$h = \frac{32 + \frac{z}{2}}{\left(2.6 + 2z + \frac{1}{z}\right)(I + .03)} \tag{12}$$

where z is expressed in microns and I in amperes e in volts and h in ohms per micron. To facilitate solving for z when h and I are given, a set of curves showing this relation is given in Fig. 3. It is this modulation factor h which measures the efficiency of the transmitter button.

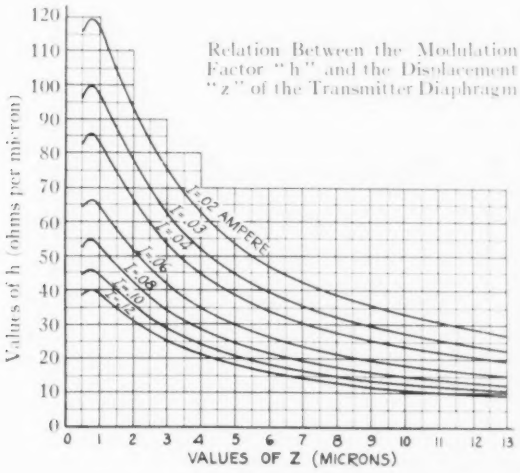


Fig. 3

It is also necessary to know the dependence of T upon z and I. To obtain this relation corresponding values of e and V, the DC drop across the transmitter as measured by direct current measuring instruments, were obtained for various degrees of agitation and amounts of direct current. Four transmitters were used in establishing the relation, the results being shown in Fig. 4. Then, for any value of the supply current I a value of T can be obtained from V. From the corresponding e a value of h and z can be obtained from equations (10) and (11). In this way the relations shown in Figs. 5

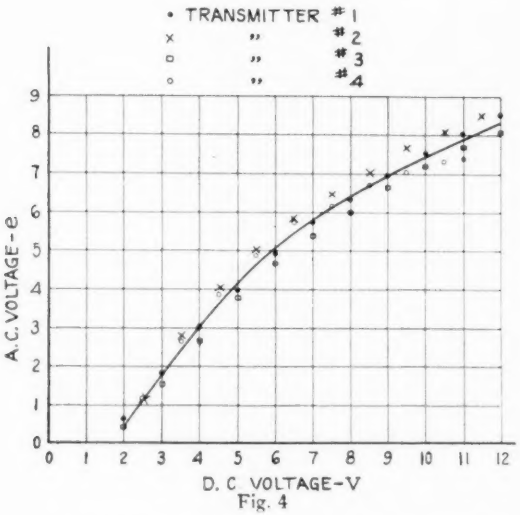

• TRANSMITTER #1
× " #2
□ " #3
○ " #4
A.C. VOLTAGE-e
D.C VOLTAGE-V

Fig. 4

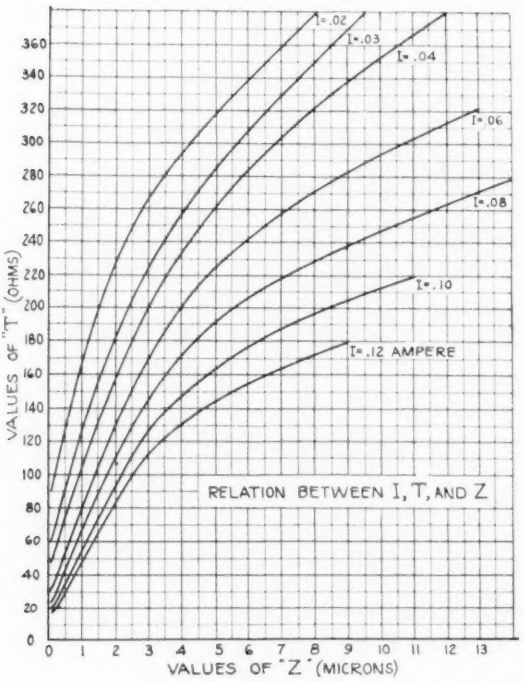

I=.02
I=.03
I=.04
I=.06
I=.08
I=.10
I=.12 AMPERE
RELATION BETWEEN I, T, AND Z
VALUES OF "T" (OHMS)
VALUES OF "Z" (MICRONS)

Fig. 5

and 6 were obtained. It is thus seen that for a given type of transmitter if the direct current and any one of the four quantities e, h, z, or T are known, the others are determined and may be obtained from suitable curves.

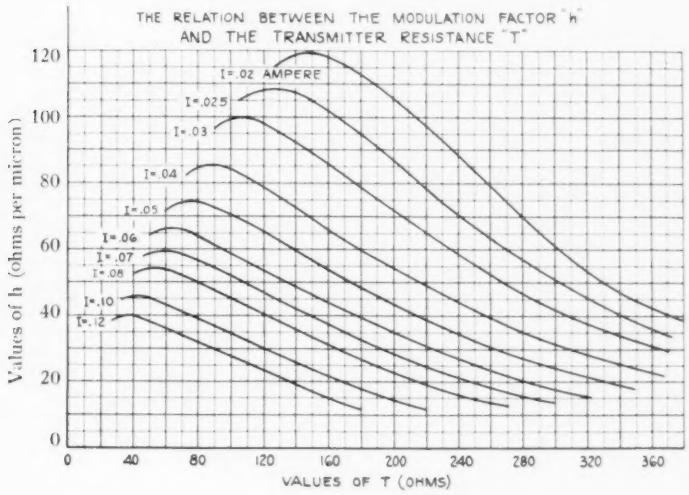


Fig. 6

Commercial receivers and transmitters have constants which vary largely from those given above. These values represent the general behavior of such instruments and are useful in understanding their operation in a howling circuit. Inasmuch as the performance of such instruments particularly the transmitter depends very largely upon the condition of operation the constants given cannot be applied with confidence to conditions greatly different from those mentioned in the paper. With these facts concerning telephone instruments in mind we are now in a position to treat some special cases.

Case 1—Diaphragms Connected Mechanically by a Rigid and Weightless Lever

To illustrate the method of solution this special case will be solved in some detail. A diagrammatic sketch illustrating the connections is shown in Fig. 7. Neglecting the reaction of the air, the vibration of the receiver diaphragm is controlled by the force Zi exerted by the

receiver winding and the opposing force X exerted by the connecting rod.

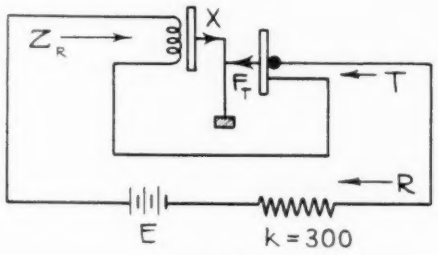


Fig. 7

The amplitude of motion of the receiver diaphragm is then given by

$$y=\frac{Zi-X}{S_R} \tag{13}$$

If the lever is rigid and weightless and has an arm ratio c, then

$$F_T=cF_R \tag{14}$$

and due to the restraint

$$y=cz=\frac{c^2X}{S_T} \tag{15}$$

Using these equations together with equation (11) it is seen that

$$M=\frac{IhZ}{cS_R+\frac{1}{c}S_T} \tag{16}$$

$$S=S_R+\frac{1}{c^2}S_T, \tag{17}$$

$$R=Z_d+j\frac{\omega Z^2}{S}+k. \tag{18}$$

The relation between I and T is given by

$$I=\frac{E}{T+R_{DC}+k} \tag{19}$$

where R_{DC} is the direct current resistance of the receiver winding and k is the line resistance. The condition (3) for howling then becomes

$$IhZ=(Z_d+k+T)\left(cS_R+\frac{1}{c}S_T\right)+j\omega Z^2c. \tag{20}$$

This is equivalent to two scalar equations and taken together with (19) and the curves of Fig. 6 gives the necessary four equations to solve for the unknowns f, h, T, and I.

The solution, however, is not straightforward since the relation between h, T, and I is only given empirically by a set of curves. By "cut and try" methods the solution for any numerical case can be obtained. The last term of (20) is usually negligible or at least it is of second order of magnitude. Consequently, the sum of the phase angles of the other factors must be approximately equal to the phase of Z. This completes the formal solution for this case.

The solution of a numerical case throws considerable light upon the physical phenomenon taking place, and also upon the method of calculation. Let the arm ratio be unity, a case corresponding to that when the diaphragms are connected directly together, and assume that the supply current is furnished by a battery of 24 volts through a line having a resistance of 300 ohms. Using the constants for the receivers and transmitters given above and expressing f in kilocycles, T in ohms, I in amperes and h in ohms per micron, equations (19) and (20) become

$$I = \frac{24}{384 + T}. \qquad (19')$$

$$Ih\ 52 \mid 24^\circ = [393 + T + 60f + j(43 + 150f)]\ [-2.14f^2 + 2.3 + j.14f] + j\ 1.7f. \qquad (20')$$

If I is positive there is no solution for f, since the angle of the first factor is in the first quadrant, and that of the second factor either in the first or second; consequently, the phases cannot match at any frequency. If the supply current is reversed, then I is negative or 180° is added to the phase of the left hand member making it a positive 156°. The solution for this case is

$f = 1072$ cycles	$i = 8.2$ mils
$h = 64$	$e = 5.5$ volts
$T = 150$ ohms	$y = z = 1.9$ microns
$I = 45$ mils	

If a value of c equal to 2.7, which is approximately equal to the square root of the ratio of mechanical impedances of the two diaphragms, then the solution for reversed DC supply becomes

$f = 1001$ cycles	$i = 10$ mils
$h = 47.3$	$e = 7.16$ volts
$T = 236$ ohms	$z = 3.9$ microns
$I = 39$ mils	$y = 10.5$ microns

It is thus seen that changing the ratio arm has increased the howling intensity, but the increase for the various elements is greatly different. The frequency is slightly lowered, the values of h and I have been

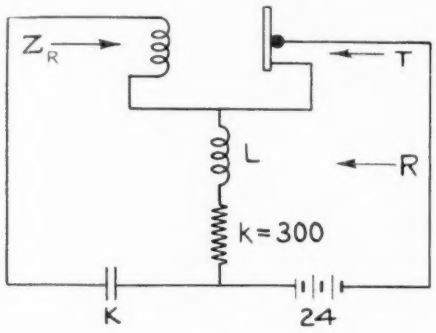


Fig. 8

reduced by 26% and 14% respectively, while the values of y, z, T, i and e have been increased 400%, 105%, 57%, 22% and 30% respectively.

If the circuit of Fig. 7 is modified as shown in Fig. 8, the inductance L being very large, then the condition for howling becomes

$$\frac{IhZ}{cS_R+\frac{1}{c}S_T} = T+Z_d+\frac{j}{K\omega}+j\frac{\omega Z^2}{S} \tag{21}$$

and

$$I = \frac{24}{300+T} \tag{22}$$

Using the same constants as above the condition for howling becomes

$$Ih\ 52\ |\overline{24°}. = \left[93+T+60f+j(43+150f)-\frac{158}{Kf}\right][-2.14f^2+2.3+j.14f]+j1.7f \tag{23}$$

The solution for values of $K=1$ *mf*, $K=1/2$ *mf*, and $K=1.5$ *mf* are given in Table I. When $K=1$ *mf* and the supply current is direct the solution which satisfies the phase equality is $f=506$. This corresponds to $h=220$ which is an impossible value. Therefore, no howling will be sustained for this condition. For $K=1.2$ *mf* the system will howl for both direct and reversed supply current, the frequency changing suddenly from 839 to 1119 cycles as the current is reversed while the other variables change only slightly.

TABLE I

	$K=1$		$K=1/2$		$K=1/5$	
	Direct	Reversed	Direct	Reversed	Direct	Reversed
f	.....	1016	839	1119	935	.....
h	220	53.5	53.5	57.4	44.2	.....
T	.....	275	160	140	220	.....
I	.....	42	52.2	54.5	46	.....
i	.....	20	18.3	17.6	10.9	.....
e	.....	8.2	6.15	5.6	7.5	.....
z	.....	5.9	2.2	1.8	3.7	.....
y	.....	16	5.95	4.9	10	.....

It is interesting to note the change in the howling frequency as the value of K increases. When the supply current is negative, and for values larger than 1 *mf*, the frequency of howling is always close to 1000, as K goes from 1 to 1/2 the frequency increases to above 1100. For smaller values of K the frequency continues to slowly increase until, for values smaller than 1/3, the system ceases to sustain oscillations. For positive values of supply current no howling will result until K becomes smaller than 2/3 where the frequency is around 800. The frequency then increases reaching a howling frequency around 1000 for $K=1/7$. For smaller values of K no howling will be sustained.

Case II—Diaphragms Coupled Together by a Small Chamber of Air

It will be assumed that the air chamber is so small that the phase of the pressure variation is the same on both diaphragms. Let V be the volume of air between the diaphragms. Then

$$V = V_0 + Q_R y + Q_T z \tag{24}$$

where V is the volume of air in the undisturbed state and Q_R and Q_T are the effective areas of the receiver and transmitter diaphragms respectively.

The pressure variation in the chamber (changes considered adiabatic) is given by

$$dp = -\gamma \frac{P}{V} dV = -(Q_R y + Q_T z)\gamma \frac{P}{V} \tag{25}$$

When the steady state is set up this may be considered a vector equation and the variables expressed in *rms* values.

The equations of motion for the diaphragms are

$$y = \frac{Zi - Q_R dp}{S_R} \tag{26}$$

and

$$z = \frac{Q_T dp}{S_T}. \tag{27}$$

Solving

$$y = \frac{Zi\left(S_T + \frac{\gamma P}{V} Q_T^2\right)}{S_R S_T + \frac{\gamma P}{V} Q_T^2 S_R + \frac{\gamma P}{V} Q_R^2 S_T}, \tag{28}$$

$$z = -\frac{\frac{\gamma P}{V} Q_R Q_T}{S_T + \frac{\gamma P}{V} Q_T^2} y, \tag{29}$$

$$M = \frac{IhZQ_R Q_T}{\frac{V}{\gamma P} S_R S_T + Q_T^2 S_R + Q_R^2 S_T}. \tag{30}$$

In this case the ratio between z and y is not fixed, but depends upon S_T which is a function of the frequency.

The loaded stiffness of the receiver diaphragm is

$$S=\frac{\frac{V}{\gamma P}S_R S_T+Q_T^2 S_R+Q_R^2 S_T}{\frac{V}{\gamma P}S_T+Q_T^2} \tag{31}$$

For the transmitter and receiver used

$$Q_R=6.5,$$
$$Q_T=10.3.$$

Let the volume of entrapped air be taken as 10 cc., then

$$\frac{\gamma P}{V}=1.418\times 10^5.$$

Using these values and the values for S_R and S_T and the circuit of Fig. 8 with $K=1/2$ the condition for howling becomes

$$Ih\,3.48\,|\,24^\circ=27.6f^5+(50.3+.459T)f^4-59.9f^3-(1.01T+85.7)f^2+31.9$$
$$+(.539T+33)+2\left[68f^5+16.7f^4-(.0506T+11.1)f^3-40f^2\right.$$
$$\left.+(.0537T-233)f+23.2+\frac{172}{f}\right] \tag{32}$$

where I is expressed in amperes, T in ohms, f in kilocycles and h in ohms per micron.

For reverse current or negative I the solution is

$f=970$ kilocycles	$i=24\ \underline{17}^\circ$
$h=30.5$	$e=8.7$ volts
$T=290$ ohms	$z=7.0$ microns
$I=.0407$ mils	$y=1.9\ \underline{158}^\circ$ microns

Comparing this to the case where the diaphragms are coupled by a lever having an arm ratio 2.7 it is seen that the air coupling produces a greater e.m.f. in the transmitter and only a slightly increased AC current. The receiver diaphragm in this case, however, has a smaller amplitude than the transmitter diaphragm. At this particular howling frequency the transmitter diaphragm stiffness is only about 1/4 that of the receiver diaphragm stiffness which explains this anomalous result. Also, it will be seen that the diaphragms vibrate almost oppositely in phase.

These cases are sufficient to illustrate the method of calculation, but there is one other important case for which I desire to give the results as this is the case handled experimentally by Kennelly and Upson.

Case III—Diaphragms Connected Acoustically by a Tube of Air of Uniform Cross-Section with an Air Chamber at Both Ends

In this case the two diaphragms are connected acoustically by the air, but since the tube has considerable length phase differences exist

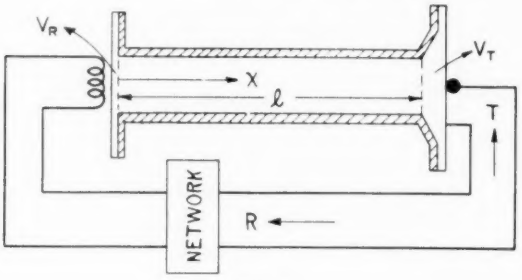


Fig. 9

at different points along it. The connections are shown schematically in Fig. 9.

The equation of motion for the receiver diaphragm is

$$y = \frac{Zi - Q_R dp_R}{S_R} = \frac{Zi}{S} \tag{33}$$

and for the transmitter diaphragm is

$$z = \frac{Q_I dp_T}{S_T} \tag{34}$$

where dp_R and dp_T are the pressure variations in the air chambers at the receiver and transmitter ends of the tube respectively.

The equations of motion for a gas in which the movements are small and in only one direction and in which the fluid friction is neglected are as follows:[6]

$$\frac{d^2\phi}{dt^2} = a^2 \frac{d^2\phi}{dx^2}, \tag{35}$$

$$\frac{dp}{\rho} = -\frac{d\phi}{dt}, \tag{36}$$

[6] See Rayleigh "Theory of Sound," Vol. II, pp. 14 and 15, 49 and 50.

where ϕ is the velocity potential, t the time, a the velocity of sound in the air, x the distance along the tube, p the pressure and ρ the density of the air.

For the case in which we are interested, a sinusoidal oscillation is sustained, so that the special solution

$$\phi = e^{j\omega t}\left(A\cos\frac{\omega X}{a} + B\sin\frac{\omega x}{a}\right) \tag{37}$$

is suitable for our problem. Quantities A and B are arbitrary constants which are determined by the end conditions. Substituting this value of ϕ in equation (35), there results

$$dp = -\rho j\omega e^{j\omega t}\left(A\cos\frac{\omega x}{a} + B\sin\frac{\omega x}{a}\right) \tag{38}$$

It remains then to determine the arbitrary constants A and B.

At the receiver end of the tube, the displacement, ζ_R of the air diaphragm across the end of the tube is related to the displacement y of the receiver diaphragm. This relationship is established by the following consideration. If q is the cross-section of the tube, the increase in volume in the air chamber is given by

$$dV_R = (\zeta_R q - yQ_R). \tag{39}$$

Assuming that the air chamber is so small that the pressure change at any instant is the same throughout, and that it takes place adiabatically, we have:

$$dp_R = -\gamma\frac{p}{V_R}dV_R \tag{40}$$

Combining equations (33), (39), and (40), we obtain:

$$qS_R\zeta_R = Q_R Zi - \left(\frac{V_R}{\gamma p}S_R + Q_R{}^2\right)dp_R \tag{41}$$

Similarly,

$$qS_T\zeta_T = \left(\frac{V_T}{\gamma p}S_T + Q_T{}^2\right)dp_T \tag{42}$$

Then the following conditions must be fulfilled at the two ends of a tube of length l.

At $x=0$, $\qquad dp = dp_R$ and $\dfrac{d\phi}{dx} = \dfrac{d\zeta_R}{dt}$;

at $x=l$, $\qquad dp = dp_T$ and $\dfrac{d\phi}{dx} = \dfrac{d\zeta_T}{dt}$.

These conditions give the following equations:

$$a\omega\rho A + S_R' B = jaZ' i_0 \tag{43}$$

$$\left(a\omega\rho \cos\frac{\omega l}{a} + S_T' \sin\frac{\omega l}{a}\right) A + \left(a\omega\rho \sin\frac{\omega l}{a} - S_T' \cos\frac{\omega l}{a}\right) B = 0 \tag{44}$$

where

$$S_R' = \frac{qS_R}{Q_R^2 + \dfrac{V_R}{\gamma p} S_R}, \quad S_T' = \frac{qS_T}{Q_T^2 + \dfrac{V_T}{\gamma p} S_T}, \quad Z' = \frac{ZQ_R}{Q_R^2 + \dfrac{V_T}{\gamma p} S_R}.$$

Solving for the constants A and B, we find their values to be:

$$A = jaZ' i_0 \left(S_T' \cos\frac{\omega l}{a} - a\omega\rho \sin\frac{\omega l}{a}\right) \div D, \tag{45}$$

$$B = jaZ' i_0 \left(S_T' \sin\frac{\omega l}{a} + a\omega\rho \cos\frac{\omega l}{a}\right) \div D, \tag{46}$$

where

$$D = [S_R' S_T' - (a\omega\rho)^2] \sin\frac{\omega l}{a} + a\omega\rho (S_R' + S_T') \cos\frac{\omega l}{a}. \tag{47}$$

The two pressure values are then given by:

$$dp_R = Z_1' a\omega\rho \left(S_T \cos\frac{\omega l}{a} - a\omega\rho \sin\frac{\omega l}{a}\right) \div D, \tag{48}$$

$$dp_T = Z_1' a\omega\rho \, S_T' \div D, \tag{49}$$

and

$$y = \frac{Zi}{S_R D}\left[S_R' S_T' - (a\omega\rho)^2 \left(1 - Q_R \frac{Z'}{Z}\right) \sin\frac{\omega l}{a} + a\omega\rho \left(S_R' S_T' \left(1 - Q_R \frac{Z'}{Z}\right)\right) \cos\frac{\omega l}{a}\right], \tag{33'}$$

$$z = \frac{qZia\omega\rho}{D}\left[\frac{Q_T}{Q_T^2 + \dfrac{V_T}{\gamma p} S_T} \; \frac{Q_R}{Q_R^2 + \dfrac{V_R}{\gamma p} S_R}\right]. \tag{34'}$$

The loaded stiffness of the receiver diaphragm is given by

$$S = \frac{qQ_T Q_R a\omega\rho \left(N \sin\dfrac{\omega l}{a} + P \cos\dfrac{\omega l}{a}\right)}{S_T \dfrac{a\omega\rho}{\gamma p}\left[\left(q^2 - \dfrac{a\omega\rho}{\gamma p} V_R V_T\right) \sin\dfrac{\omega l}{a} + q(V_T + V_R) \cos\dfrac{\omega l}{a} - a\omega\rho \left(\dfrac{a\omega\rho}{\gamma p} V_R Q_R^2 \sin\dfrac{\omega l}{a} + qQ_T \cos\dfrac{\omega l}{a}\right)\right.}, \tag{50}$$

where

$$N = S_T S_R \frac{1}{a\omega\rho}\left[\frac{q^2}{Q_R Q_T} - \frac{(a\omega\rho)^2}{(\gamma p)^2}\frac{V_T V_R}{Q_R Q_T}\right] - S_R\left(\frac{a\omega\rho}{\gamma p}\right)\frac{Q_T}{Q_R}V_R - S_T\left(\frac{a\omega\rho}{\gamma p}\frac{Q_R}{Q_T}\right)V_T, \quad (51)$$

$$P = S_R\frac{Q_T}{Q_R} + S_T\frac{Q_R}{Q_T} + S_T S_R\frac{V_T + V_R}{Q_T Q_R}\frac{1}{\gamma p}. \quad (52)$$

The unilateral mutual impedance M is given by

$$M = \frac{IhZ}{N\sin\frac{\omega l}{a} + P\cos\frac{\omega l}{a}} \quad (53)$$

The condition for sustained howling becomes

$$\frac{IZ}{T+R}h = N\sin\frac{\omega l}{a} + P\cos\frac{\omega l}{a}. \quad (54)$$

If the two diaphragms work directly into the connecting tube as pistons, then $Q_R = Q_T = q = Q$ and $V_R = V_T = O$ and the expressions for M and S become [7]

$$M = \frac{IhZ\,Q\,a\omega\rho}{[S_R S_T - (a\omega\rho)^2 Q^2]\sin\frac{\omega l}{a} + (S_R + S_T)Q\,a\omega\rho\cos\frac{\omega l}{a}}, \quad (55)$$

$$S = \frac{[S_R S_T - (a\omega Q)^2]\sin\frac{\omega l}{a} + (a\omega\rho Q)(S_R + S_T)\cos\frac{\omega l}{a}}{S_T\sin\frac{\omega l}{a} + a\omega\rho\,Q\cos\frac{\omega l}{a}}. \quad (56)$$

The method of solution is the same as that given for the simpler cases, although it is evident that the actual work of calculation is more involved.

It is seen that in such a system the intensity and frequency depend upon a large number of quantities, namely: S_T and S_R, the diaphragm stiffness factors; Q_R and Q_T the effective areas of the two diaphragms; V_R and V_T the volumes of air entrapped between the diaphragm and the opening into connection tube; the length l and the cross section q of the connecting tube; the pressure a, the density s, and the velocity of sound a for the gas in the connecting tube; the resistance T, direct current I and modulation factor h of the transmitter; and

[7] These two equations were given by H. W. Nichols in essentially this form in the Physical Review, Vol. 10, p. 171; 1917.

the force factor and impedance of the receiving circuit. Modification of any of these may produce marked changes in the resulting howling.

The way the length l enters the formula (54) for sustained howling indicates that the curves representing the possible frequencies of howling, that is, frequencies which produce equality of phase on both sides of the equation, vary periodically with the length.

The intersection of the branches of these curves on any given frequency line will be separated by distances corresponding to $\frac{a}{f}$, that is, corresponding to a wave length at the pitch corresponding to f. Also, if the supply current is reversed, that is, the sign of I

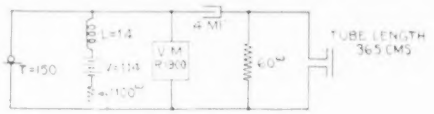


Fig. 10

changed, and the length of the tube varied until the frequency of howling is brought back to the original value, the change in length must be equal to $\frac{a}{2f}$. For since the frequency is unchanged all the quantities in equation (54) remain unchanged except the sine and cosine factors. Adding a half wave length is equivalent to adding π to the angle which makes the left hand member the negative of its first value, and consequently, restores the phase equality.

Using the circuit shown in Fig. 10 for the electrical coupling, the frequency of howling was computed for various tube lengths, the results being given in Fig. 11.

The instrument constants were those used before, the other values being $V_R = 1.6$ cc., $V_T = 6.4$ cc., and $q = .97$ cm.^{2}, $a = 3.43 \times 10^4$ cm/sec. $\rho = .001203$ gm/cm^3. Using these values the formulae for N and P become

$$N = (-1.31f^5 + 7.5f^3 - 9.68f + 3.26\frac{1}{f}) \times 10^8 + j(.141f^4 - .63f^2 + .36) \times 10^8,$$

$$P = (5.5f^4 - 12.35f^2 + 6.77) \times 10^8 + j(-.60f^3 + .66f) \times 10^8,$$

where f is the frequency in kilocycles.

The points on the calculated curves of Fig. 11 were obtained by direct experimental observation with the circuit shown, and with various lengths of brass tube coupling the transmitter and receiver together. The agreement between the calculated and observed

values is well within the experimental error involved in determining the constants used in the calculation.

In Fig. 12 are shown similar calculated curves for a transmitter called "hollow," that is, for one having a lower natural period of vibration. It is coupled to the same receiver as used before. The dotted curves in each case represent the behavior for reversed current.

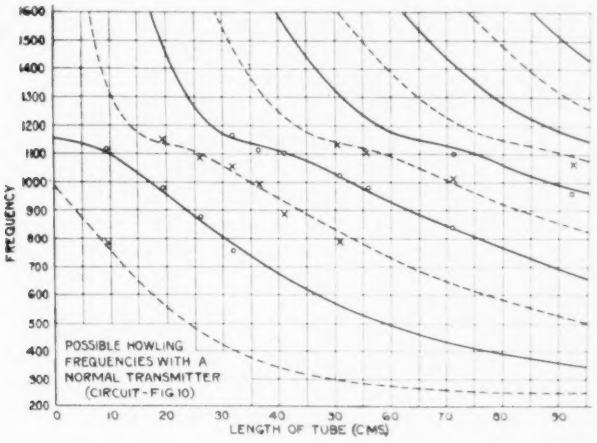


Fig. 11

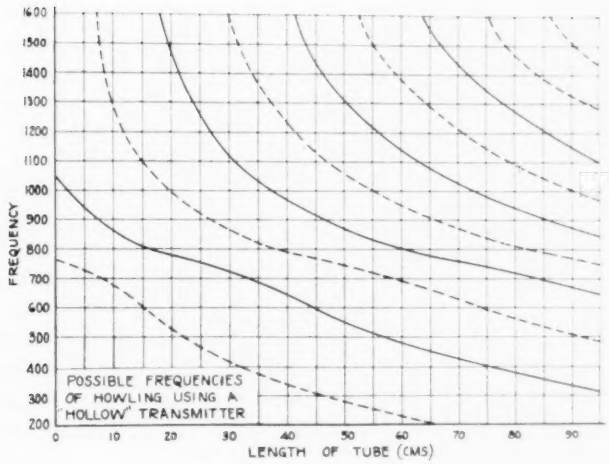


Fig. 12

In Figs. 13 and 14 are shown the probable frequencies of howling for these two transmitters as the tube length of the coupler is increased. The shaded areas are the so-called breaking points where the howling may be at either of the frequencies shown.

With these facts in mind let us review the conclusions reached by Kennelly and Upson given in the beginning of this paper. It is seen that conclusion (1) is not warranted. The transmitter and circuit conditions as well as the receiver diaphragm influence the mean frequency of humming. The second conclusion regarding the branches

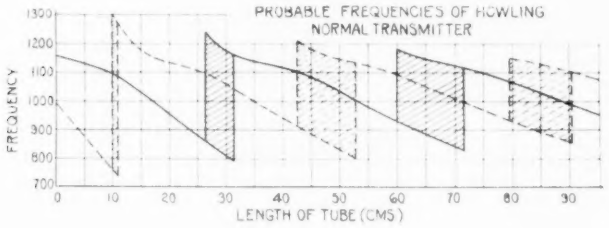


Fig. 13

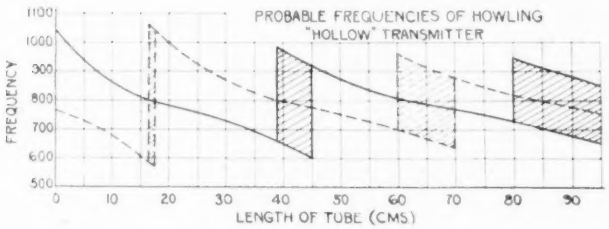


Fig. 14

of the curves representing the relation between frequency and tube length is correct and the explanation has just been given. This periodic relation is not only true of the mean frequency line but for every constant frequency line.

The terms corresponding to $\frac{1}{4}\frac{V}{n_o}$ and $\frac{3}{4}\frac{V}{n_o}$ depend upon a number of factors including the circuit and end conditions. Conclusion (3) is partially correct, the range of the howling frequencies depending upon the efficiencies of the transmitter, receiver, and circuit is evident from equation (54). Calculations show that conclusion (4) is generally correct although not necessarily so.

When the transmitter and receiver are coupled by the air in an open room the behavior is somewhat similar to the case just solved. The size and shape of the room as well as the disposition of articles of furniture will all influence the intensity and frequency of howling. In general when the two instruments are moved apart the frequency will go up and down similar to that when they are coupled by a tube.

NOMENCLATURE

T	Transmitter Resistance.
R	Impedance looking away from Transmitter Terminals.
Z_R	Impedance of Receiver.
Z_A	Damped Impedance of Receiver.
e	Electromotive Force Created in the Transmitter.
i	Alternating Current in the Transmitter Branch.
M	Unilateral Mutual Impedance between Receiver Current and Transmitter e.m.f.
F_R	Force on Receiver Diaphragm.
F_T	Force on Transmitter Diaphragm.
S_R	Stiffness Factor of Receiver Diaphragm.
S_T	Stiffness Factor of Transmitter Diaphragm.
y	Receiver Diaphragm Displacement.
z	Transmitter Diaphragm Displacement.
m	Mass of Diaphragm.
r	Mechanical Resistance of Diaphragm.
s	Elastic Constant of Diaphragm.
f	Frequency.
ω	2π times Frequency.
Z	Force Factor of Receiver.
S	Loaded Stiffness of Receiver Diaphragm.
I	Direct Current Supplied to Transmitter.
V	DC Voltage Drop across Transmitter Terminals.
h	Modulation Factor of the Transmitter.
N	Mechanical Force on Receiver Diaphragm for Case I.
E	Electromotive Force of Supply Battery.
k	Resistance in Line for Case I.
K	Capacity of Condenser.
V_R	Volume of Air in Front of Receiver Diaphragm.
V_T	Volume of Air in Front of Transmitter Diaphragm.
Q_R	Effective Area of Receiver Diaphragm.
Q_T	Effective Area of Transmitter Diaphragm.
p	Air Pressure.
γ	Adiabatic Constant.
ϕ	Velocity of Potential.
a	Velocity of Sound in Air.
x	Distance Along Connecting Tube.
ρ	Density of Air.
ξ_R	Displacement of Air Particle at Receiver End of Tube.
ξ_T	Displacement of Air Particle at Transmitter End of Tube

Electric Circuit Theory and the Operational Calculus[1]

By JOHN R. CARSON

CHAPTER VI

Propagation of Current and Voltage Along the Non-Inductive Cable

THE principal practical applications of the operational calculus in electrotechnics are to the theory of the propagation of current and voltage along transmission systems. Of such transmission systems the simplest is the non-inductive cable. The theory of the non-inductive cable is not only of great historic interest, relating as it does to Kelvin's early work on the possibility of transatlantic telegraphy, but is also of very considerable practical importance today, and serves as a basis for the theory of submarine telegraphy over long distances. We shall therefore consider the propagation phenomena in the non-inductive cable in some detail.

The propagation phenomena in any type of transmission system are isolated and exhibited in the clearest possible manner when we confine attention to the infinitely long line, with voltage applied directly to the line terminals. Furthermore, as we shall see later, the solution for the infinitely long line is fundamental and can be extended to the more practical case of the finite line with terminal impedances. We therefore, in this chapter, shall confine our attention to the case of the infinitely long cable with voltage applied directly to the cable terminals.

Consider a cable of distributed resistance R and capacity C per unit length, extending from $x=0$ along the positive x axis. From a previous chapter (see equations (64) and (65)), we are in possession of the operational equations of voltage and current; they are, for the infinitely long line,

$$V=e^{-\sqrt{\alpha p}}\ V_0, \tag{162}$$

$$I=\frac{1}{Rx}\sqrt{\alpha p}\ e^{-\sqrt{\alpha p}}\ V_0=\sqrt{\frac{Cp}{R}}e^{-\sqrt{\alpha p}}\ V_0, \tag{163}$$

where $\alpha=x^2RC$, and V_0 is the terminal cable voltage at $x=0$. Let us now assume that the terminal voltage V_0 is a "unit e.m.f."; then

$$V=e^{-\sqrt{\alpha p}}, \tag{164}$$

$$I=\frac{1}{Rx}\sqrt{\alpha p}\ e^{-\sqrt{\alpha p}}. \tag{165}$$

[1] Continued from the October, 1925, issue.

The solution of (164) for V was considered in some detail in the preceding chapter; it is, by (129)

$$V=\frac{1}{\sqrt{\pi}}\int_0^\tau \frac{e^{-1/\tau}}{\tau\sqrt{\tau}}\,d\tau \tag{166}$$

where $\tau=4t/\alpha=4t/x^2RC$. Series expansions of this solution were also given. Another equivalent form is, by (131)

$$V=1-\frac{2}{\sqrt{\pi}}\int_0^{1/\sqrt{\tau}} e^{-\tau^2}\,d\tau. \tag{167}$$

This last form, recognizable also from inspection of the series expansion (132), is useful because the integral term is what is called the error function and has been completely computed and tabulated.

Before discussing these formulas and the light they throw on propagation phenomena in the non-inductive cable, we shall derive the solution for the current. A very simple way of doing this is to make use of the differential equation (57)

$$I=-\frac{1}{R}\frac{\partial}{\partial x}V.$$

Now from (166) and the relation

$$\frac{\partial}{\partial x}=\frac{d\tau}{dx}\frac{d}{d\tau}$$

we get

$$\frac{\partial}{\partial x}V=\frac{1}{\sqrt{\pi}}\frac{e^{-1/\tau}}{\tau\sqrt{\tau}}\frac{d}{dx}\frac{4t}{x^2RC}$$

$$=-\frac{2}{x\sqrt{\pi}}\frac{e^{-1/\tau}}{\sqrt{\tau}},$$

whence

$$I=\frac{2}{xR\sqrt{\pi}}\frac{e^{-1/\tau}}{\sqrt{\tau}}=\sqrt{\frac{C}{\pi Rt}}e^{-1/\tau}. \tag{168}$$

It is worthwhile verifying the formula by direct solution from the operational equation (165). From formula (g) of the table of integrals, we have

$$h=e^{-2\sqrt{\lambda p}}\sqrt{p}\sqrt{\frac{C}{R}}$$

$$=\frac{e^{-\lambda/t}}{\sqrt{\pi}t}\sqrt{\frac{C}{R}}.$$

Comparison with the operational equation shows that they are identical, within a constant factor provided we put $\lambda = \alpha/4$. Consequently the solution of (165) is

$$I = \sqrt{\frac{C}{\pi R t}}\, e^{-\alpha/4t} = \sqrt{\frac{C}{\pi R t}}\, e^{-1/\tau}$$

which agrees with (168). This, it may be remarked, is an excellent example of the utility of the table of integrals in solving operational equations.

This formula is easily calculated for large values of t by expanding the exponential function; it is

$$\frac{2}{Rx}\,\frac{1}{\sqrt{\pi\tau}}\left[1-\left(\frac{1}{\tau}\right)+\frac{1}{2!}\left(\frac{1}{\tau}\right)^2-\cdots\right].$$

The propagation phenomena of the non-inductive cable are therefore determined by the pair of equations

$$V = \frac{1}{\sqrt{\pi}}\int_0^{\tau}\frac{e^{-1/\tau}}{\tau\sqrt{\tau}}\,d\tau = 1-\frac{2}{\sqrt{\pi}}\int_0^{1/\sqrt{\tau}} e^{-\tau^2}\,d\tau \tag{169}$$

and

$$I = \frac{2}{\sqrt{\pi x R}}\,\frac{e^{-1/\tau}}{\sqrt{\tau}} = \sqrt{\frac{C}{\pi R t}}\, e^{-1/\tau} \tag{170}$$

where $\tau = 4t/\alpha = \dfrac{4t}{x^2RC}$.

Now an important feature of these formulas is that the voltage at point x is a function only of $\dfrac{4}{x^2RC}\,t$; that is, of *4t divided by the total resistance and capacity* of the cable from $x=0$ to $x=x$. The same statement holds for the *form* of the current wave: its magnitude, however, is inversely proportional to xR, or the total resistance of the cable up to point x. Consequently a single curve, with proper time scale serves to give the voltage wave at any point on the cable. Similarly a single curve, with proper time and amplitude scales, serves to depict the current wave at any distance from the cable terminals. These curves are given in Figs. 3 and 4.

Referring to the curve depicting the current wave, we observe that it is finite for all values of $t>0$; consequently, in the ideal cable, the velocity of propagation is infinite. This is a consequence, of course, of the fact that the distributed inductance of the cable is neglected. Actually, of course, the velocity of propagation cannot exceed the

velocity of light. The error, however, in neglecting the inductance in the case of long cables is appreciable only near the head of the wave provided we confine attention to d.c. or low frequency voltages. This point will be discussed and explained more fully in connection with the transmission line.

The current, while finite, is negligibly small until τ reaches the

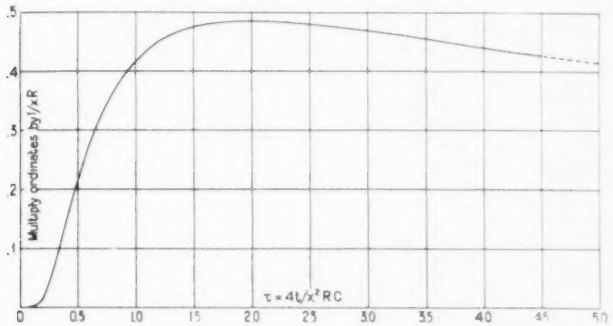


Fig. 3—Current in non-inductive cable ($G=0$) unit e.m.f. applied

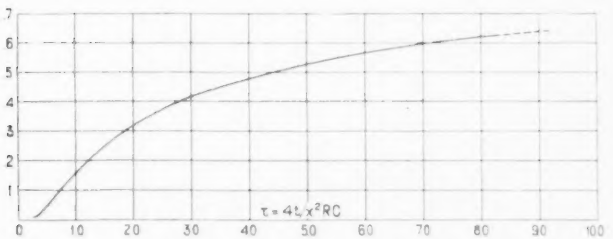


Fig. 4—Voltage in non-inductive cable ($G=0$) unit e.m.f. applied

value 0.2. In the neighborhood of this point it begins to build up rapidly; reaches at $\tau=2$ its maximum value

$$\frac{2}{\sqrt{\pi}\,xR}\,\frac{e^{-0.5}}{\sqrt{2}}=\frac{2}{\sqrt{\pi}\,xR}(0.429)$$

and then begins to decrease, ultimately dying away in accordance with the formula

$$\frac{2}{\sqrt{\pi}\,xR}\,\frac{1}{\sqrt{\tau}}\left(1-\frac{1}{\tau}+\frac{1}{2!}\left(\frac{1}{\tau}\right)^2-\cdots\right).$$

Its subsidence to its final zero value is very slow; for example, when $\tau = 100$ its value is still

$$\frac{2}{\sqrt{\pi}\, xR}(0.10).$$

Turning to the voltage curve, Fig. 4, we see that it is negligibly small until τ reaches the value 0.25, at which point it begins to build up. Its maximum rate of building up occurs when $\tau = 2/3$, after

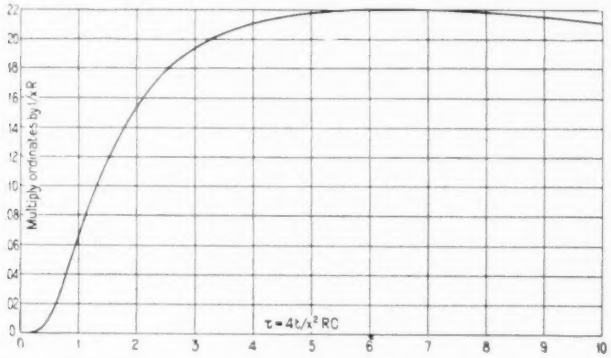


Fig. 5—Power transmitted in non-inductive cable ($G = 0$)

which it builds up more and more slowly. Its approach to its final steady value is in accordance with the formula

$$V = 1 - \frac{2}{\sqrt{\pi\tau}}\left(1 - \frac{1}{3\tau} + \frac{1}{2!}\frac{1}{5\tau^2} - \cdots\right).$$

Even, therefore, when τ is as great as 100, V differs sensibly from its ultimate value, unity, its value being 0.8876.

Since the actual time is $\frac{x^2RC}{4}\tau$, it follows that the speed of building up is inversely proportional to the square of the length of the cable.

The power curve VI is given in Fig. 5. V.I is the rate at which energy is being transmitted past the point x of the cable.

The fact that the form of the current and voltage waves depends only on $4t/x^2RC$ is at the basis of Kelvin's famous "KR" law, long applied to cable telegraphy and sometimes incorrectly applied to telephony. When the first transatlantic telegraph cable was under consideration, Kelvin attacked the problem of propagation along the non-inductive cable and arrived at formulas equivalent to (169) and

(170). From these formulas he announced the law that the "speed" of the cable, i.e., the number of signals transmissible per unit time, is inversely proportional to the product of the total capacity and total resistance of the cable (KR in the English notation). To see just what this means requires a little digression into the elementary theory of telegraph transmission.

Telegraph signals are transmitted in code by means of "dots" and "dashes." The "dot" is the signal which results when a battery is impressed on the cable for a definite interval of time, after which the cable is short circuited. A "dash" is the same except that the time interval during which the battery is connected to the cable is increased. The "dots" and "dashes" are separated by intervals, called "spaces", during which the cable is short circuited. Now when the cable is short-circuited we may imagine a negative battery impressed on the cable in series with the original battery. Consequently the current in the cable, corresponding to a signal composed of a series of dots, dashes and spaces, will be represented by a series of the form

$$I(t)-I(t-t_1)+I(t-t_2)-I(t-t_3)+I(t-t_4)-\ \ldots \qquad (171)$$

where, in the cable under consideration, $I(t)$ is given by (168). t_1 is the duration of the first impulse, t_2-t_1 of the first space, t_3-t_2 of the second impulse, etc.

Now by (168)

$$I(t)=\frac{2}{xR\sqrt{\pi}}\frac{e^{-1/\tau}}{\sqrt{\tau}}=\frac{2}{xR\sqrt{\pi}}\phi(\tau).$$

τ is, of course, $4t/x^2CR=4t/KR$ (in the English notation). Now suppose that

$$\tau_1=\frac{4t_1}{x^2CR},$$

$$\tau_2=\frac{4t_2}{x^2CR}, \text{ etc.}$$

Then the signal can be written as

$$\frac{2}{xR\sqrt{\pi}}\{\phi(\tau)-\phi(\tau-\tau_1)+\phi(\tau-\tau_2)-\ \ldots\ \} \qquad (172)$$

Now if the relative time intervals τ_1, τ_2 . . . are kept constant (as the length of the cable is varied), the actual time intervals t_1, t_2 . . . are proportional to x^2CR or to KR, and the wave form of the total signal is independent of KR, when referred to the relative time scale τ.

Hence, if T is the total time of the signal, T is proportional to x^2CR (or to KR). That is to say, if the duration of the component dots, dashes, spaces of the signal are proportional to the "KR" of the cable, the wave form of the received signal, referred to the τ time scale, is invariable, and the total time required to transmit the signal is proportional to the "KR" of the cable. Now the maximum theoretical speed of transmission on the cable is limited by the requirement that the received signal shall bear a recognizable likeness to the original system of dots and dashes: in other words there is a

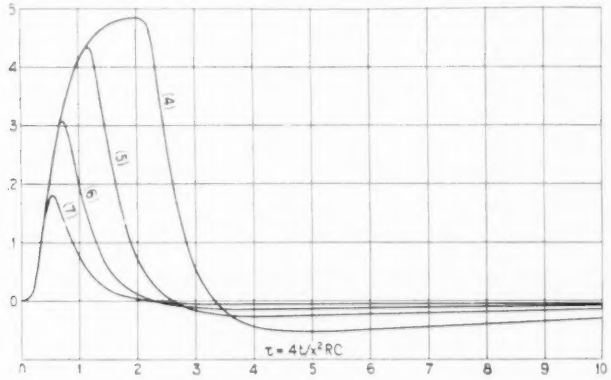


Fig. 6—Elementary telegraph signals in non-inductive cable

maximum allowable departure in wave form between received and transmitted signals. If, therefore, the actual speeds of two cables are inversely proportional to their "KRs," the wave form will be the same. This establishes Kelvin's "KR" law. As a corollary, if the length of the cable is doubled the speed of signaling is reduced to one-quarter, assuming the same definition of signals.

The foregoing will be somewhat clearer, perhaps, if we refer to curves, 4, 5, 6, 7 of Fig. 6 which illustrate the distortion suffered by elementary dot signals in cable transmission. Curve 4 shows the dot signal produced by a unit battery applied to the cable terminals for a time interval $t=2\,\frac{x^2RC}{4}$, while curves 5, 6 and 7 are the corresponding dot signals when the battery is applied for the time intervals $\frac{x^2RC}{4}$, $\frac{1}{2}\,\frac{x^2RC}{4}$ and $\frac{1}{4}\,\frac{x^2RC}{4}$. Any further decrease in the duration of the impressed dot, beyond that shown in curve 7, does not

affect the *shape* of the transmitted dot, which means that the cable speed has reached its theoretical maximum. These curves, it should be observed, can be interpreted in two ways. First, we can regard the length x of the cable as fixed and the duration of the impressed dot as varied. On the other hand, we can regard the actual duration of the impressed dot as constant and the length of the cable as varied. From the latter standpoint the curves illustrate the progressive distortion of the signal as it is transmitted along the cable.

The dot signal of relative duration T can be written as

$$\begin{aligned} D &= I(\tau), \qquad \tau < T \\ &= I(\tau) - I(\tau - T), \qquad \tau < T \end{aligned}$$

and the second expression can be expanded in a Taylor's series, giving

$$D = T\frac{d}{d\tau}I(\tau) - \frac{T^2}{2!}\frac{d^2}{d\tau^2}I(\tau) + \cdots$$

If T is sufficiently short this becomes

$$D = T\ I'(\tau). \tag{173}$$

Hence when the dot signal is of sufficiently short relative duration T, the wave shape of the received signal is constant, $I'(\tau)$, and its amplitude is proportional to the relative duration of the dot.

This can be generalized for any type of transmission system: Let the dot signal be produced by an e.m.f. $f(t)$ of actual duration T. Then the received dot signal, by formula (31), is

$$\begin{aligned} D &= \frac{d}{dt}\int_0^t f(\tau)I(t-\tau)d\tau, \qquad t < T \\ &= \frac{d}{dt}\int_0^T f(\tau)I(t-\tau)d\tau, \qquad t > T. \end{aligned}$$

For $t > T$ this becomes

$$D = I'(t)\int_0^T f(\tau)d\tau - I''(t)\int_0^T \tau f(\tau)d\tau + \cdots$$

and for sufficiently short duration T, we have approximately,

$$D = I'(t)\int_0^T f(\tau)d\tau. \tag{174}$$

Hence for a sufficiently short duration of the impressed e.m.f. the received dot signal is of constant wave form, independent of the shape of the impressed e.m.f., and its amplitude is proportional to the time

integral of the impressed e.m.f. These principles are of considerable practical importance in telegraphy.

The *leaky cable*, that is, a cable with distributed leakage conductance G in addition to resistance R and capacity C, is of some interest. The differential equations of the problem are given in equations (70); the operational formulas for the case of voltage directly impressed on the terminals of the infinitely long line are

$$V = e^{-x\sqrt{CRp+RG}}\, V_o,$$

$$I = \sqrt{\frac{pC}{R} + \frac{G}{R}}\, e^{-x\sqrt{CRp+RG}}\, V_o.$$

Writing $CRx^2 = \alpha$ and $RGx^2 = \beta$, $G/C = \lambda$, and assuming a "unit e.m.f." impressed on the cable, this becomes

$$V = e^{-\sqrt{\alpha p+\beta}}, \tag{175}$$

$$I = \sqrt{\frac{C}{R}}\sqrt{p+\lambda}\; e^{-\sqrt{\alpha p+\beta}}. \tag{176}$$

These equations are readily solved by means of the table and formulas given in a preceding chapter.

But first let us attempt to solve the operational equation (175) for the voltage by Heaviside methods, guided by the solution of the operational equation

$$V = e^{-\sqrt{\alpha p}} \tag{124}$$

of the preceding chapter. Expand the exponential function in (175) in the usual power series; it is

$$V = 1 - \sqrt{\alpha p+\beta} + \frac{(\alpha p+\beta)}{2!} - \frac{(\alpha p+\beta)\sqrt{\alpha p+\beta}}{3!} + \cdots \tag{177}$$

Now discard the integral terms and write

$$V = 1 - \left\{1 + \frac{\alpha p+\beta}{3!} + \frac{(\alpha p+\beta)^2}{5!} + \cdots\right\}\sqrt{\alpha p+\beta}. \tag{178}$$

We have now to interpret the expression $\sqrt{\alpha p+\beta}$. We have by ordinary algebra

$$\begin{aligned}\sqrt{\alpha p+\beta} &= \left(1+\frac{\beta}{\alpha p}\right)^{1/2}\sqrt{\alpha p} = \left(1+\frac{\lambda}{p}\right)^{1/2}\sqrt{\alpha p} \\ &= \left[1+\frac{\lambda}{2p} - \frac{1}{2!}\left(\frac{\lambda}{2p}\right)^2 + \frac{1.3}{3!}\left(\frac{\lambda}{2p}\right)^3 + \cdots\right]\sqrt{\alpha p}.\end{aligned} \tag{179}$$

Now identify $\sqrt{p}$ with $1/\sqrt{\pi t}$ in accordance with the Heaviside rule, and $1/p$ with $\int dt$. We get

$$\sqrt{\alpha p+\beta}=\sqrt{\frac{\alpha}{\pi t}}\left\{1+\frac{\lambda t}{1!}-\frac{(\lambda t)^2}{3!}+\frac{1.4}{5!}(\lambda t)^3-\ldots\right\}. \tag{180}$$

Now in the terms of the expansion (178) identify p^n with d^n/dt^n and substitute (180); we get

$$\begin{aligned} V=1-\Bigg\{1+\frac{1}{3!}\left(\alpha\frac{d}{dt}+\beta\right)+\frac{1}{5!}\left(\alpha^2\frac{d^2}{dt^2}+2\alpha\beta\frac{d}{dt}+\beta^2\right)+\ldots\Bigg\} \\ \times\sqrt{\frac{\alpha}{\pi t}}\left\{1+\frac{\lambda t}{1!}-\frac{(\lambda t)^2}{3!}+1.4\frac{(\lambda t)^3}{5!}-\ldots\right\}. \end{aligned} \tag{181}$$

This series is hopelessly complicated to either interpret or compute. It is, in fact, an excellent illustration of the grave disadvantages under which many of Heaviside's series solutions labor. We shall therefore attack the solution by aid of the theorems and formulas of a preceding section. The simplicity of the solutions which result is remarkable.

The operational formula for the voltage is

$$V=e^{-\sqrt{\alpha p+\beta}}. \tag{175}$$

Now the operational formula for the voltage in the non-leaky cable is (see equation (164))

$$V=e^{-\sqrt{\alpha p}}.$$

In order to distinguish between the two cases, let us denote the voltage in the latter case by V^0; thus

$$V^0=e^{-\sqrt{\alpha p}}. \tag{182}$$

Now by theorem (VII) and equation (182) we have

$$\begin{aligned} V^0e^{-\lambda t}&=\frac{p}{p+\lambda}\,e^{-\sqrt{\alpha(p+\lambda)}}, \\ &=\frac{p}{p+\lambda}\,e^{-\sqrt{\alpha p+\beta}}. \end{aligned} \tag{183}$$

Now write (175) as

$$\begin{aligned} V&=\frac{p+\lambda}{p}\cdot\frac{p}{p+\lambda}\,e^{-\sqrt{\alpha p+\beta}}, \\ &=\left(1+\frac{\lambda}{p}\right)\cdot\frac{p}{p+\lambda}\,e^{-\sqrt{\alpha p+\beta}}. \end{aligned} \tag{184}$$

It follows at once by comparison with (183) and the rule that $1/p$ is to be replaced by $\int dt$, that

$$V = \left(1 + \lambda \int_0^t dt\right) V^0 e^{-\lambda t}. \tag{185}$$

By a precisely similar procedure with the operational formula (176) for the current, we get

$$I = \left(1 + \lambda \int_0^t dt\right) I^0 e^{-\lambda t} \tag{186}$$

where I^0 is the current in the non-leaky cable. Now by formulas (169) and (170)

$$V^0 = \frac{1}{\sqrt{\pi}} \int_0^{4L^2\alpha} \frac{e^{-1/t}}{t\sqrt{t}}\, dt, \tag{169}$$

$$I^0 = \sqrt{\frac{C}{\pi R t}}\, e^{-\alpha^2 L^2 t}, \tag{170}$$

which completes the formal solution of the problem.

Formulas (185) and (186) are extremely interesting, first as showing the superiority of the definite integral to the series expansion—compare (185) with the series expansions (181)—and secondly as exhibiting clearly the effect of leakage on the propagated waves of current and voltage. We see that in both the current and voltage the effect of leakage is two-fold: first it attenuates the wave by the factor $e^{-\lambda t}$, $(\lambda = G/C)$, and secondly it adds a component consisting of the progressive integral of the attenuated wave. This, it may be remarked, is the general effect of leakage in all types of transmission systems. Its effect is, therefore, easily computed and interpreted.

Formulas (185) and (186) are very easy to compute with the aid of a planimeter or integraph; or, failing these devices, by numerical integration. However, for large values of t, the character of the waves is more clearly exhibited if we make use of the identity

$$\int_0^t dt = \int_0^\infty dt - \int_t^\infty dt$$

whence

$$V = \left(1 + \lambda \int_0^\infty dt\right) V^0 e^{-\lambda t} - \lambda \int_t^\infty V^0 e^{-\lambda t} dt \tag{187}$$

and

$$I = \left(1 + \lambda \int_0^\infty dt\right) I^0 e^{-\lambda t} - \lambda \int_t^\infty I^0 e^{-\lambda t} dt. \tag{188}$$

The first two terms of these formulas are clearly the ultimate steady state values of the voltage and current waves, and can be determined by evaluating the infinite integrals. A far simpler and more direct way, however, is to make use of the fact that the ultimate steady values of V and I are gotten from the operational formulas by setting $p=0$. That this statement is true is easily seen if we reflect that the steady d.c. voltage and current are gotten from the original differential equations of the problem by assuming a steady state and setting $d/dt=0$.

From the operational formulas we get, therefore,

$$\left(1+\lambda\int_0^\infty dt\right)V''e^{-\lambda t}=e^{-x\sqrt{\lambda}}=e^{-x\sqrt{RG}}, \tag{189}$$

$$\left(1+\lambda\int_0^\infty dt\right)I''e^{-\lambda t}=\sqrt{\frac{C\lambda}{R}}\,e^{-x\sqrt{\lambda}}=\sqrt{\frac{G}{R}}\,e^{-x\sqrt{RG}}. \tag{190}$$

Introducing these expressions into (187) and (188) respectively, we get

$$V=e^{-x\sqrt{RG}}-\lambda\int_t^\infty V''e^{-\lambda t}dt, \tag{191}$$

$$I=\sqrt{\frac{G}{R}}\,e^{-x\sqrt{RG}}-\lambda\int_t^\infty I''e^{-\lambda t}dt. \tag{192}$$

The definite integrals can be expanded by partial integration; thus

$$-\lambda\int_t^\infty V''e^{-\lambda t}dt=\int_t^\infty V''de^{-\lambda t}$$

$$=-V''e^{-\lambda t}-\int_t^\infty e^{-\lambda t}\frac{d}{dt}V''dt.$$

Continuing this process we get

$$V=e^{-x\sqrt{RG}}-e^{-\lambda t}\left(1+\frac{d}{\lambda dt}+\frac{d^2}{\lambda^2dt^2}+\cdots\right)V'', \tag{193}$$

$$I=\sqrt{\frac{G}{R}}e^{-x\sqrt{RG}}-e^{-\lambda t}\left(1+\frac{d}{\lambda dt}+\frac{d^2}{\lambda^2dt^2}+\cdots\right)I''. \tag{194}$$

Using the values of V'' and I'', as given by (169) and (170), it is extremely easy to compute V and I, for large values of t, from (193) and (194).

So far we have considered the current and voltage waves in response to a "unit e.m.f.," impressed on the cable at $x=0$. It is of interest and importance to examine the waves due to sinusoidal e.m.fs., suddenly impressed on the cable, particularly in view of proposals to employ alternating currents in cable telegraphy.

We start with the fundamental formula

$$x(t)=\frac{d}{dt}\int_0^t f(t-\tau)h(\tau)d\tau$$
$$=\int_0^t f(t-\tau)h'(\tau)d\tau$$

provided $h(o)=0$, which is the case in the cable.

If $f(t)=\sin \omega t$, we write

$$x_s(t)=\sin \omega t\int_0^t \cos \omega t.h'(t)dt$$
$$-\cos \omega t\int_0^t \sin \omega t.h'(t)dt. \tag{194-a}$$

Similarly, if the impressed e.m.f. is $\cos \omega t$,

$$x_c(t)=\cos \omega t\int_0^t \cos \omega t.h'(t)dt$$
$$+\sin \omega t\int_0^t \sin \omega t.h'(t)dt. \tag{194-b}$$

The investigation of the building-up of alternating currents and voltages, therefore, depends on the progressive integrals

$$C=\int_0^t \cos \omega t.h'(t)dt,$$
$$S=\int_0^t \sin \omega t.h'(t)dt. \tag{194-c}$$

For the case of the *voltage* waves on the non-inductive, non-leaky cable these integrals, by aid of equations (169), become, if we write $\omega'=\alpha\omega/4$,

$$C=\frac{1}{\sqrt{\pi}}\int_0^\tau \frac{e^{-1/\tau}\cos \omega'\tau}{\tau\sqrt{\tau}}d\tau,$$
$$S=\frac{1}{\sqrt{\pi}}\int_0^\tau \frac{e^{-1/\tau}\sin \omega'\tau}{\tau\sqrt{\tau}}d\tau, \tag{194-d}$$

where, as before, $\tau=4t/\alpha$.

For the current wave we have, by (170),

$$C=\frac{2}{\sqrt{\pi}\,xR}\int_0^\tau \left(\frac{1}{\tau}-\frac{1}{2}\right)\frac{e^{-1/\tau}\cos \omega'\tau}{\tau\sqrt{\tau}}d\tau,$$
$$S=\frac{2}{\sqrt{\pi}\,xR}\int_0^\tau \left(\frac{1}{\tau}-\frac{1}{2}\right)\frac{e^{-1/\tau}\sin \omega'\tau}{\tau\sqrt{\tau}}d\tau. \tag{194-e}$$

For small values of τ and ω' these integrals can be numerically evaluated without great labor. Mechanical devices, such as the Coradi Harmonic Analyzer, are here of great assistance. In fact the Coradi Analyzer gives these progressive integrals automatically. It may be said, therefore, that a complete mathematical investigation of the building-up of alternating current and voltage waves on the non-inductive cable presents no serious difficulties, although the labor of computation is necessarily considerable. One fact makes the complete investigations much less laborious than might be sup-

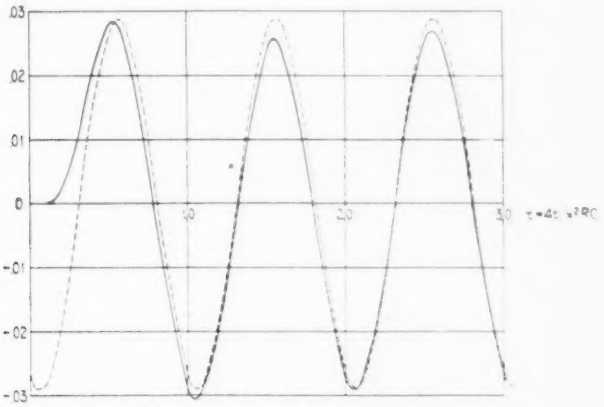


Fig. 7.—Non-inductive cable ($G=0$), building-up of alternating current.
Applied e.m.f. $\cos \omega t$; $\omega = 2\pi \frac{4}{x^2RC}$

posed. This is, if the foregoing integrals are calculated for a given value of ω', the results apply to all lengths of cable and all actual frequencies $\omega/2\pi$, such that $\alpha\omega$ is a constant. Then if we double the length of the cable and quarter the frequency, the integrals are unaffected.

The solid curve of Fig. 7 shows the building-up of the cable voltage in response to an e.m.f. $\cos \omega t$, impressed at time $t=0$. The frequency $\omega/2\pi$ is so chosen that $\omega' = \alpha\omega/4 = 2\pi$, and the curve is calculated from equations (194-b) and (194-e). The dotted curve shows the corresponding *steady-state* voltage on the cable; that is, the voltage which would exist if the e.m.f. $\cos \omega t$ had been applied at a long time preceding $t=0$. We observe that, for this frequency, the building-up is effectually accomplished in about one cycle, and that the transient distortion is only appreciable during the first half-cycle.

The case is very much different when a higher frequency is applied. Fig. 8 shows the building-up of the alternating current in the cable when an e.m.f. sin ωt is applied at time $t=0$. The frequency is so chosen that $\omega' = \alpha\omega/4 = 10\pi$. The outstanding features of this curve are that the initial current surge is very large compared with the final steady-state, and that the transient distortion is relatively very large. It is evident that the frequency here shown could not be

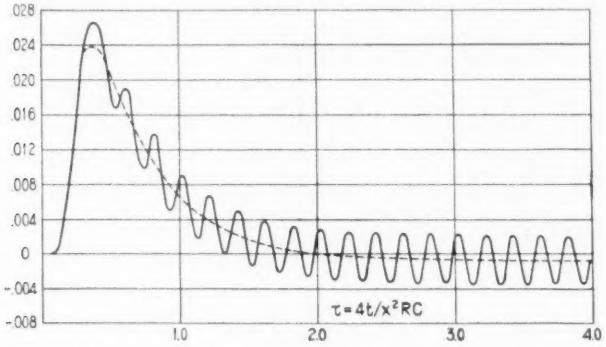


Fig. 8—Non-inductive cable ($G=0$). Building-up of alternating current.

Applied e.m.f. sin ωt: $\omega = 10\pi \dfrac{4}{x^2RC}$

employed for signaling purposes. This curve has been computed from the steady-state formulas, and equations (160) and (161) for the transient distortion.

If the applied frequency $\omega/2\pi$ is very high, the steady-state becomes negligibly small, and the complete current is obtained to a good approximation by taking the leading terms of (160) and (161). Thus if the applied e.m.f. is sin ωt, and ω is sufficiently large, the cable current is

$$\frac{2}{\sqrt{\pi x R}} \frac{1}{\omega'} \frac{d}{d\tau} \frac{e^{-1/\tau}}{\sqrt{\tau}}$$

by (160) and (170) while, if the impressed e.m.f. is cos ωt, it is

$$\frac{2}{\sqrt{\pi x R}} \left(\frac{1}{\omega'}\right)^2 \frac{d^2}{d\tau^2} \frac{e^{-1/\tau}}{\sqrt{\tau}}$$

by (161) and (170). Here $\omega' = \alpha\omega/4$ and $\tau = 4t/\alpha$.

CHAPTER VII

The Propagation of Current and Voltage Along the Transmission Line

We now take up the more important and difficult problem of investigating the propagation phenomena in the transmission line. The transmission line has distributed series resistance R and inductance L, and distributed shunt capacity C and leakage conductance G. It is the addition of the series inductance L which makes our problem more difficult and at the same time introduces the phenomena of true propagation with finite velocity, as distinguished from the diffusion phenomena of the cable problem. The cable theory serves very well for the problems of trans-oceanic telegraphy [8] but is quite inadequate in the problems of telephonic transmission.

If I denotes the current and V the voltage at point x on the line, the well known differential equations of the problem are:—

$$\left(L\frac{d}{dt}+R\right) I = -\frac{\partial}{\partial x} V,$$
$$\left(C\frac{d}{dt}+G\right) V = -\frac{\partial}{\partial x} I. \tag{195}$$

Replacing d/dt by p, these become

$$(Lp+R)\, I = -\frac{\partial}{\partial x} V,$$
$$(Cp+G)\, V = -\frac{\partial}{\partial x} I. \tag{196}$$

From the second of these equations

$$\frac{\partial V}{\partial x} = -\frac{1}{Cp+G}\frac{\partial^2}{\partial x^2} I$$

and substitution in the first gives

$$(Lp+R)(Cp+G)\, I = \frac{\partial^2}{\partial x^2} I. \tag{197}$$

Similarly if we eliminate I, we get

$$(Lp+R)(Cp+G)\, V = \frac{\partial^2}{\partial x^2} V. \tag{198}$$

[8] With the installation of the new submarine cable, continuously loaded with permalloy, this statement must be modified. In this cable, the inductance plays a very important part, and is responsible for the greatly increased speed of signaling obtainable.

If we assume a solution of the form

$$V = Ae^{-\gamma x} + Be^{\gamma x}$$

where A and B are arbitrary constants, substitution shows that the solution satisfies the differential equation for V provided

$$\gamma^2 = (Lp+R)(Cp+G). \tag{199}$$

From equation (196) it then follows that

$$\begin{aligned} I &= \frac{\gamma}{Lp+R}(Ae^{-\gamma x} - Be^{\gamma x}) \\ &= \frac{Cp+G}{\gamma}(Ae^{-\gamma x} - Be^{\gamma x}). \end{aligned} \tag{200}$$

Now restricting attention to the infinitely long line extending along the positive x axis, with voltage V_0 impressed directly on the line at $x=0$, the reflected wave vanishes and we get

$$\begin{aligned} V &= V_0 e^{-\gamma x}, \\ I &= \frac{Cp+G}{\gamma} V_0 e^{-\gamma x}, \\ \gamma^2 &= (Lp+R)(Cp+G). \end{aligned} \tag{201}$$

Now let us write

$$\gamma^2 = \frac{1}{v^2}[(p+\rho)^2 - \sigma^2] \tag{202}$$

where

$$\begin{aligned} v &= 1/\sqrt{LC}, \\ \rho &= \frac{R}{2L} + \frac{G}{2C}, \\ \sigma &= \frac{R}{2L} - \frac{G}{2C}. \end{aligned}$$

Then setting $V_0 = 1$, the *operational equations* of the problem become

$$V = e^{-\frac{x}{v}\sqrt{(p+\rho)^2-\sigma^2}}, \tag{203}$$

$$I = v\left(C + \frac{G}{p}\right) p \frac{e^{-\frac{x}{v}\sqrt{(p+\rho)^2-\sigma^2}}}{\sqrt{(p+\rho)^2-\sigma^2}}. \tag{204}$$

Now consider the operational equation, defining a new variable F:

$$F = p \frac{e^{-\frac{x}{v}\sqrt{(p+\rho)^2-\sigma^2}}}{\sqrt{(p+\rho)^2-\sigma^2}}. \tag{205}$$

It follows at once from our operational rules, and (203) and (204), that

$$I = v\left(C + G\int_0^t dt\right)F, \tag{206}$$

$$V = -v\int_0^t \frac{\partial F}{\partial x}dt. \tag{207}$$

Our problem is thus reduced to evaluating the function F, from the operational equation (205). This equation can be solved by aid of the operational rules and formulas already given. The process is rather complicated, and there is less chance of error if we deal instead with the integral equation of the problem

$$\frac{e^{-\frac{x}{v}\sqrt{(p+\rho)^2-\sigma^2}}}{\sqrt{(p+\rho)^2-\sigma^2}} = \int_0^\infty F(t)e^{-pt}dt. \tag{208}$$

Now let us search through our table of definite integrals. We do not find this integral equation as it stands, but we do observe that formula (m) resembles it, and this resemblance suggests that formula (m) can be suitably transformed to give the solution of (208). We therefore start with the formula

$$\frac{e^{-\lambda\sqrt{p^2+1}}}{\sqrt{p^2+1}} = \int_\lambda^\infty e^{-pt}J_0(\sqrt{t^2-\lambda^2})\,dt. \tag{m}$$

This, regarded as an integral equation, defines a function which is zero for $t<\lambda$ and has the value $J_0(\sqrt{t^2-\lambda^2})$ for $t\geqq\lambda$, J_0 being the Bessel function of order zero. We now transform (m) as follows:

(1) Let $\lambda p = q$ and $t/\lambda = t_1$. Substituting in (m) we get

$$\frac{e^{-\sqrt{q^2+\lambda^2}}}{\sqrt{q^2+\lambda^2}} = \int_1^\infty e^{-qt_1}J_0(\lambda\sqrt{t_1^2-1})\,dt_1.$$

Now, in order to keep our original notation in p and t, replace q by p and t_1 by t; we get

$$\frac{e^{-\sqrt{p^2+\lambda^2}}}{\sqrt{p^2+\lambda^2}} = \int_1^\infty e^{-pt}J_0(\lambda\sqrt{t^2-1})\,dt. \tag{m.1}$$

(2) In (m.1) make the substitution $p=q+\mu$ and then in the final expression replace q by p; we get

$$\int_1^\infty e^{-pt}e^{-\mu t}J_0(\lambda\sqrt{t^2-1})dt = \frac{e^{-\sqrt{(p+\mu)^2+\lambda^2}}}{\sqrt{(p+\mu)^2+\lambda^2}}. \tag{m.2}$$

(3) In (m.2) make the substitution $p=\frac{x}{v}q$ and $t_2=\frac{x}{v}t$, and ultimately replace q by p and t_2 by t; we get

$$\int_{x/v}^{\infty} e^{-pt}.e^{-\mu_1 t}J_o\left(\lambda_1\sqrt{t^2-\frac{x^2}{v^2}}\right)dt=\frac{e^{-\frac{x}{v}\sqrt{(p+\mu_1)^2+\lambda_1^2}}}{\sqrt{(p+\mu_1)^2+\lambda_1^2}} \tag{m.3}$$

where $\lambda_1=\frac{v}{x}\lambda$ and $\mu_1=\frac{v}{x}\mu$. (They are, of course, as yet, arbitrary parameters, except that they are restricted to positive values).

(4) Now if we compare (m.3) with the integral equation (208) for F, we see that they are identical provided we get

$$\mu_1=\rho,$$
$$\lambda_1=i\sigma=\sigma\sqrt{-1},$$

which is possible, since $\rho>\sigma$.

Introducing these relations, we have

$$\int_{x/v}^{\infty} e^{-pt}.e^{-\rho t}\, I_o(\sigma\sqrt{t^2-x^2/v^2})dt=\frac{e^{-\frac{x}{v}\sqrt{(p+\rho)^2-\sigma^2}}}{\sqrt{(p+\rho)^2-\sigma^2}}. \tag{m.4}$$

Here I_o denotes the Bessel function of imaginary argument; thus $J_o(iz)=I_o(z)$.

It follows from (m.4) and the integral equation (208) that

$$\begin{aligned} F(t)&=0 \text{ for } t<x/v, \\ &=e^{-\rho t}I_o(\sigma\sqrt{t^2-x^2/v^2}) \text{ for } t\geqq x/v. \end{aligned} \tag{209}$$

Having now solved for $F=F(t)$, the current and voltage are gotten from equations (206) and (207). Thus

$$\begin{aligned} I&=0 \text{ for } t<x/v, \\ &=\sqrt{\frac{C}{L}}\,F(t)+vG\int_{x/v}^{t}F(t)dt \text{ for } t\geqq x/v. \end{aligned} \tag{210}$$

The corresponding voltage formula is

$$\begin{aligned} V&=0 \text{ for } t<x/v, \\ &=e^{-\rho x/v}+\frac{\sigma x}{v}\int_{x/v}^{t}\frac{e^{-\rho\tau}I_1(\sigma\sqrt{\tau^2-x^2/v^2})}{\sqrt{\tau^2-x^2/v^2}}\,d\tau \text{ for } t\geqq x/v. \end{aligned} \tag{211}$$

Here $I_1(\sigma\sqrt{\tau^2-x^2/v^2})$ is the Bessel function of order 1: thus $-iJ_1(iz)=I_1(z)$. The function is entirely real. The derivation of formula (211) is a little troublesome, owing to the discontinuous character of the function F: the detailed steps are given in an appendix.

The preceding solution depends for its outstanding directness and simplicity on the recognition of the infinite integral identity (m), into which the integral equation of the problem can be transformed. When such identities are known their value in connection with the solution of operational equations requires no emphasis. On the other hand, we cannot always expect to find such an identity in the case of every operational equation; and, particularly in the case of such an important case as the transmission equation it would be unfortunate to have no alternative mode of solution. Fortunately a quite direct series expansion solution is obtainable from the operational equation, and this will now be derived. As a matter of convenience we shall restrict the derivation to the voltage formula

$$V = e^{-\frac{x}{v}\sqrt{(p+\rho)^2-\sigma^2}}. \tag{203}$$

As a further matter of mere convenience we shall assume that $G=0$, so that $\sigma=\rho$ and (203) becomes

$$V = e^{-\tau\sqrt{p^2+2\rho p}} \tag{203-a}$$

where $\tau = x/v$.

The method holds equally well for the current equation (204) and for the general case $\sigma \neq \rho$.

Write (203-a) as

$$V = e^{-\tau p(1+2\rho/p)^{1/2}}$$

and expand the exponential factor $(1+2\rho/p)^{1/2}$ by the binomial theorem; thus

$$(1+2\rho/p)^{1/2} = 1+\frac{\rho}{p}+\alpha_2\left(\frac{\rho}{p}\right)^2+\alpha_3\left(\frac{\rho}{p}\right)^3+\ldots$$

so that

$$V = e^{-\tau p}.e^{-\rho\tau}.\exp\left(-\frac{\alpha_2\tau\rho^2}{p}-\frac{\alpha_3\tau\rho^3}{p^2}-\frac{\alpha_4\tau\rho^4}{p^3}-\ldots\right).$$

Now the operational equation

$$v = \exp\left(-\frac{\alpha_2\tau\rho^2}{p}-\frac{\alpha_3\tau\rho^3}{p^2}-\frac{\alpha_4\tau\rho^4}{p^3}-\ldots\right)$$

can be expanded in inverse powers of p; thus

$$v = 1+\frac{\beta_1}{p}+\frac{\beta_2}{p^2}+\frac{\beta_3}{p^3}+\ldots\ldots$$

the power series solution of which is

$$v(t) = 1+\frac{\beta_1 t}{1!}+\frac{\beta_2 t^2}{2!}+\frac{\beta_3 t^3}{3!}+\ldots\ldots$$

It follows at once from the preceding and Theorem VII that

$$V(t) = 0 \text{ for } t < \tau$$

$$= e^{-\rho\tau}\left(1 + \beta_1 \frac{(t-\tau)}{1!} + \beta_2 \frac{(t-\tau)^2}{2!} + \ldots\right) \text{ for } t > \tau$$

If the coefficients $\beta_1, \beta_2 \ldots$ are evaluated, a simple matter of elementary algebra, the foregoing expansion in the retarded time $t-\tau$ will be found to agree with the solution (211) when σ is put equal to ρ.

We shall now discuss the outstanding features of the propagation phenomena in the light of equations (210) and (211) for the current and voltage. We observe, first, that we have a true finite velocity of propagation $v = 1/\sqrt{LC}$. No matter what the form of impressed e.m.f. at the beginning of the line ($x = 0$), its effect does not reach the point x of the line until a time $t = x/v$ has elapsed. Consequently $v = x/t$ is the velocity with which the wave is propagated. This is a strict consequence of the distributed inductance and capacity of the line and depends only on them, since $v = 1/\sqrt{LC}$. It will be recalled that in the case of the cable, where the inductance is ignored, no finite velocity of propagation exists.

The question of velocity of propagation of the wave has been the subject of considerable confusion and misinterpretation when dealing with the steady-state phenomena. It seems worth while to briefly touch on this in passing.

As has been pointed out in preceding chapters, the symbolic or complex steady-state formula is gotten from the operational equation by replacing the symbol p by $i\omega$ where $i = \sqrt{-1}$ and $\omega/2\pi$ is the frequency. If this is done in the operational equation (203) for the voltage, the symbolic formula is

$$V = e^{-\frac{x}{v}\sqrt{(i\omega+\rho)^2 - \sigma^2}} e^{i\omega t}.$$

If the expression $\sqrt{(i\omega+\rho)^2 - \sigma^2}$ is separated into its real and imaginary parts we get an expression of the form

$$V = e^{-\alpha x} e^{i\omega\left(t - \frac{\beta x}{v}\right)},$$

where

$$\beta = \sqrt{\frac{\omega^2 + \sigma^2 - \rho^2 + \sqrt{(\omega^2 + \sigma^2 - \rho^2)^2 + 4\omega^2\rho^2}}{2\omega^2}}$$

and

$$\alpha = \rho/\beta v.$$

Now if we keep the expression $t - \beta \frac{x}{v}$ constant, that is, if we move along the line with velocity $dx/dt = v/\beta$, the phase of the wave will remain constant. This is interpreted often as meaning that the

velocity of propagation of the wave is v/β. Now since β is greater than unity and only approaches unity as the frequency becomes indefinitely great, the inference is frequently made that the velocity of propagation depends upon and increases to a limiting value v, with the frequency. This velocity, however, is not the true velocity of propagation, which is always v, but is the *velocity of phase propagation in the steady-state*. This distinction is quite important and failure to bear it in mind has led to serious mistakes.

Returning to equation (211) and (210) we see that after a time interval $t=x/v$ has elapsed since the unit e.m.f. was impressed on the cable, the voltage at point x suddenly jumps from zero to the value $e^{-\rho x/v}$ while the current correspondingly jumps to the value $\sqrt{\frac{C}{L}}\,e^{-\rho x/v}$. The exponential factor $\rho x/v$ is

$$x\left(\frac{R}{2L}+\frac{G}{2C}\right)\sqrt{LC}=x\left(\frac{R}{2}\sqrt{\frac{C}{L}}+\frac{G}{2}\sqrt{\frac{L}{C}}\right)=ax$$

which will be recognized as the *steady-state attenuation factor* for high frequencies. Similarly $\sqrt{C/L}$ is the steady-state admittance of the line for high frequencies. The sudden jumps in the current and voltage at time $t=x/v$ are called the heads of the current and voltage waves. If, instead of a unit e.m.f., a voltage $f(t)$ is impressed on the line at time $t=0$, the corresponding heads of the waves are $f(o)e^{-ax}$ and $\sqrt{C/L}\,f(o)e^{-ax}$ for voltage and current respectively. These expressions follow at once from the integral formula

$$\begin{aligned}x(t)&=\frac{d}{dt}\int_0^t f(t-\tau)h(\tau)d\tau\\&=f(o)h(t)+\int_0^t f'(t-\tau)h(\tau)d\tau.\end{aligned}$$

The tails of the waves, that is, the parts of the waves subsequent to the time $t=x/v$, are more complicated and will depend on the distance x along the line and on the line parameters ρ and σ. The two simplest cases are the *non-dissipative* line, and the *distortionless* line.

The ideal non-dissipative line, quite unrealizable in practice, is one in which both R and G are zero. In this case $\rho=\sigma=0$, and formulas (210) and (211) become

$$\begin{aligned}I&=0 \text{ for } t<x/v,\\&=\sqrt{\frac{C}{L}} \text{ for } t\geq x/v,\\V&=0 \text{ for } t<x/v,\\&=1 \text{ for } t\geq x/v.\end{aligned}$$

Both current and voltage jump, at time $t=x/v$, to their steady values. If an e.m.f. $f(t)$ is impressed on the line at time $t=0$, the corresponding current and voltage waves are

$$I=0 \text{ for } t<x/v,$$
$$=\sqrt{\frac{C}{L}}f(t-x/v) \text{ for } t\geq x/v,$$
$$V=0 \text{ for } t<x/v,$$
$$=f(t-x/v) \text{ for } t\geq x/v.$$

Consequently the ideal non-dissipative line transmits the waves with finite velocity v, without attenuation or distortion. Such a line is, of course, the ideal transmission system.

The non-dissipative line is, of course, purely theoretical and unrealizable in practice; the *distortionless* line is, however, approximately realizable, and as the name implies, transmits without distortion of wave form. The distortionless line is one in which the line constants are so related that

$$\sigma=\frac{R}{2L}-\frac{G}{2C}=0.$$

If this condition is satisfied, formulas (210) and (211) become

$$I=0 \text{ for } t<x/v,$$
$$=\sqrt{\frac{C}{L}}e^{-ax} \text{ for } t\geq x/v,$$
$$V=0 \text{ for } t<x/v,$$
$$=e^{-ax} \text{ for } t\geq x/v.$$

Furthermore, if the impressed e.m.f. is $f(t)$, the corresponding current and voltage waves are:—

$$I=0 \text{ for } t<x/v,$$
$$=\sqrt{\frac{C}{L}}e^{-ax}f(t-x/v) \text{ for } t\geq x/v,$$
$$V=0 \text{ for } t<x/v,$$
$$=e^{-ax}f(t-x/v) \text{ for } t\geq x/v.$$

The distortionless line, therefore, transmits the waves without distortion of wave form, but attenuates the waves by the factor e^{-ax}. Such a line is an ideal transmission system as regards preservation of wave form, but introduces serious attenuation losses. For example, if a line has normally negligible leakage, and leakage is introduced

to secure the condition $R/L=G/C$, the line is thereby rendered distortionless but the attenuation is doubled.

One of Heaviside's most important contributions to wire transmission theory was to point out the properties of the distortionless line, its approximately realizable character, and to base on it a correct theory of telephonic transmission.

The character of the wave propagation when the parameters ρ and σ are not restricted to special values, can only be roughly inferred from inspection of the formulas, and then only when the properties of the Bessel function I_0 and I_1 have been studied. Fortunately these functions have been computed and tabulated for small values of the argument, and have simple asymptotic expansions for large values. It is therefore a simple matter to compute and graph a representative set of curves which show the current and voltage waves for various values of ρ, σ and x. For this purpose it is convenient to introduce a change of variables and write:

$$\tau = vt$$
$$a = \rho/v$$
$$b = \sigma/v$$

whence the formulas for current and voltage become:

$$I = \sqrt{\frac{C}{L}}\, e^{-a\tau} I_0(b\sqrt{\tau^2-x^2}) + (a-b)\sqrt{\frac{C}{L}} \int_x^\tau e^{-a\tau} I_0(b\sqrt{\tau^2-x^2})\, d\tau, \tag{210a}$$

$$V = e^{-ax} + bx \int_x^\tau \frac{e^{-a\tau} I_1(b\sqrt{\tau^2-x^2})}{\sqrt{\tau^2-x^2}}\, d\tau. \tag{211a}$$

Figs. (9) to (18) give a representative set of curves illustrating the form of the propagated current and voltage waves for different lengths of line, and different values of the line parameters a and b, or ρ and σ.

The curves of Figs. (9) and (10) show the current entering the line in response to a unit e.m.f. applied at time $t=0$. The line is assumed to be non-leaky ($b=0$) and is computed for two different values of the parameter a. We see that the current instantly jumps to the value $\sqrt{C/L}$ and then begins to die away, the rate at which it dies away depending on and increasing with the parameter $a=\frac{R}{2}\sqrt{\frac{C}{L}}$.

If we now consider a point x out on the line, the current is zero until $\tau=x$, at which time it jumps to the value $\sqrt{C/L}\, e^{-ax}$. It then

begins to die away provided x and a are such that $ax<2$. If, however, we are considering a point at which $ax>2$, the current begins to rise instead of fall after the initial jump, and may attain a maximum value very large compared with the head before it starts to die away. This is shown in the curves of Figs. (11), (12) and (13), also computed

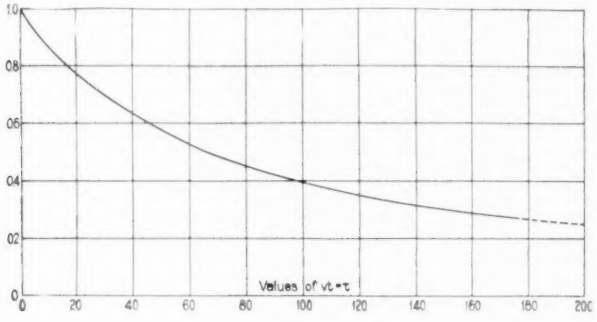


Fig. 9—Current entering line; $\frac{R}{2}\sqrt{\frac{C}{L}}=a=0.0132$; $G=0$.

Multiply ordinates by $\sqrt{C/L}$

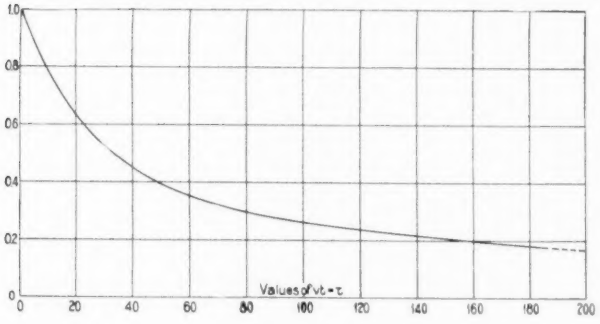


Fig. 10—Current entering line; $\frac{R}{2}\sqrt{\frac{C}{L}}=a=0.2645$; $G=0$.

Multiply ordinates by $\sqrt{C/L}$

for the non-leaky line ($b=0$). From these curves we see that, as the length of the line and the parameter a increase, the relative magnitude of the tail, as compared with the head of the wave, increases. Finally when the line becomes very long, the head of the wave becomes negligibly small, and the wave, except in the neighborhood of its head, becomes very close to that of the corresponding non-inductive

cable. This is shown in curves (13) and (14), for the line and the corresponding cable, which are plotted to the same time scale and ordinate scale to facilitate comparison. Curve (15) shows the effect of leakage in eliminating the tail. This line is not quite distortionless but nearly so.

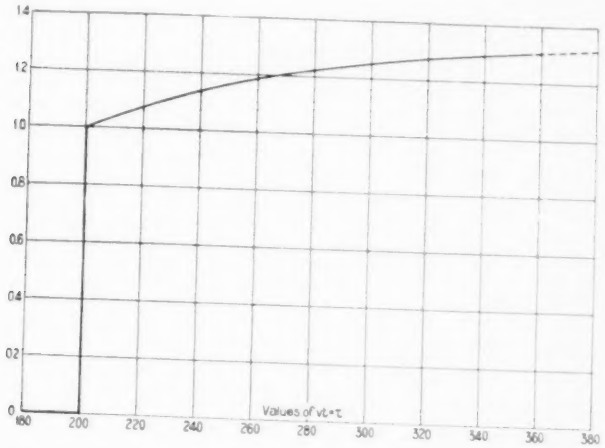


Fig. 11—Propagated current in line; $x = 200$; $\frac{R}{2}\sqrt{\frac{C}{L}} = a = 0.0132$; $G = 0$.

Multiply ordinates by $\sqrt{C/L}.e^{2.64}$

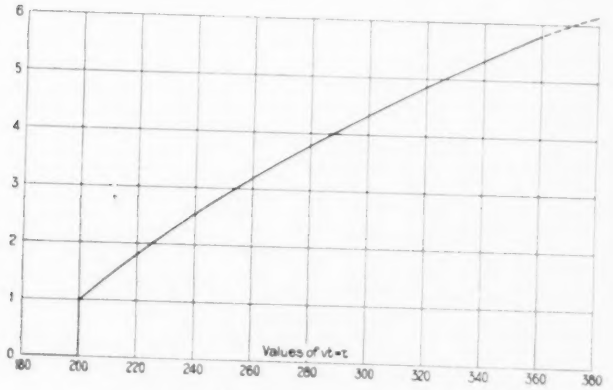


Fig. 12—Propagated current in line; $x = 200$; $\frac{R}{2}\sqrt{\frac{C}{L}} = 0.02645$; $G = 0$.

Multiply ordinates by $\sqrt{C/L}.e^{5.29}$

An interesting feature of both current and voltage waves is that when a sufficient time has elapsed after the arrival of the head of the wave, the waves become closer and closer to the wave of the corresponding non-inductive cable; that is, to the cable having the same R,C and G. Consequently the inductance plays no part in the subsidence of the waves to their final values.

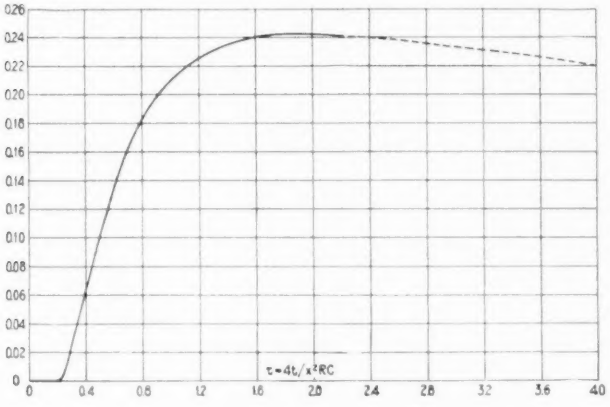


Fig. 13—Propagated current in line; $\frac{R}{2}\sqrt{\frac{C}{L}}\,x=10$; $G=0$.
Multiply ordinates by $2/Rx$.

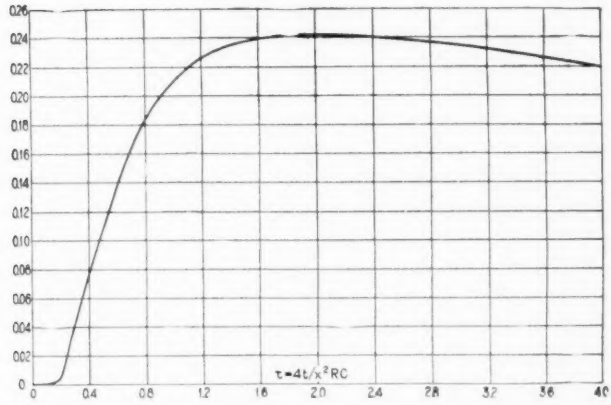


Fig. 14—Propagated current in cable. Multiply ordinates by $2/Rx$.

Curves (16), (17) and (18) illustrate the voltage wave for several conditions. After the arrival of the head, the wave slowly builds up to its final value. Curve (18) represents the case where the line is very nearly distortionless, showing how completely the distorting tail of the wave is eliminated.

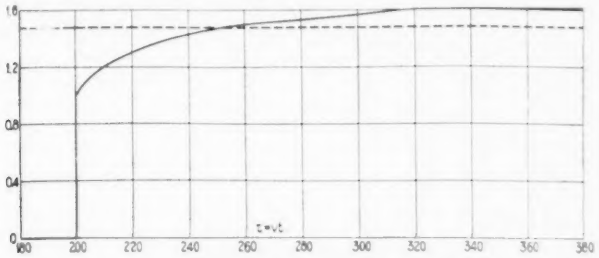


Fig. 15—Propagated current in line; $x=200$

$$a=\frac{R}{2}\sqrt{\frac{C}{L}}+\frac{G}{2}\sqrt{\frac{L}{C}}=0.0353$$

$$b=\frac{R}{2}\sqrt{\frac{C}{L}}-\frac{G}{2}\sqrt{\frac{L}{C}}=0.01765$$

Multiply ordinates by $\sqrt{C/L}e^{-7.06}$

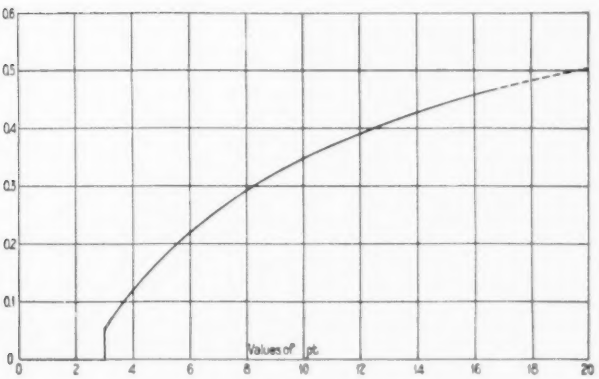


Fig. 16—Propagated voltage in line; $\frac{R}{2}\sqrt{\frac{C}{L}}\,x=ax=3;\ G=0.$

So far we have confined attention to the current and voltage waves in response to a unit e.m.f. applied at time $t=0$ to the line terminals. Of much greater technical importance is the question of the waves in response to a sinusoidal e.m.f. suddenly applied to the line termi-

nals. In order to investigate this important problem it is convenient to divide the expressions for the current and voltage waves as given by equations (210-a) and (211-a) into two components. We write for $\tau \geq x$,

$$I = \sqrt{\frac{C}{L}} e^{-ax} + J(t), \tag{210-b}$$

$$V = e^{-ax} + W(t), \tag{211-b}$$

where, by definition, $J(t)$ and $W(t)$ are the differences between the total waves and their heads. The advantage of analyzing the waves into these components is that the distortion of the waves is due to

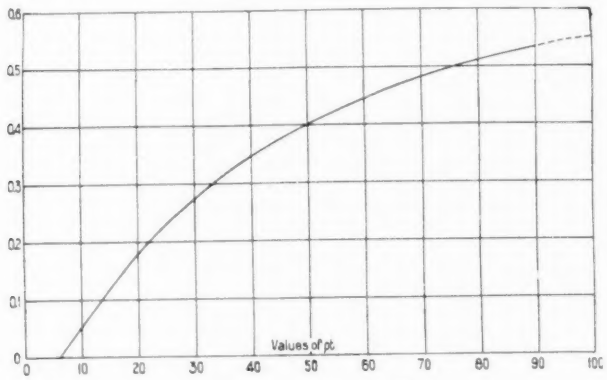


Fig. 17—Propagated voltage in line: $\frac{R}{2}\sqrt{\frac{C}{L}}\, x = ax = 6;\ G = 0.$

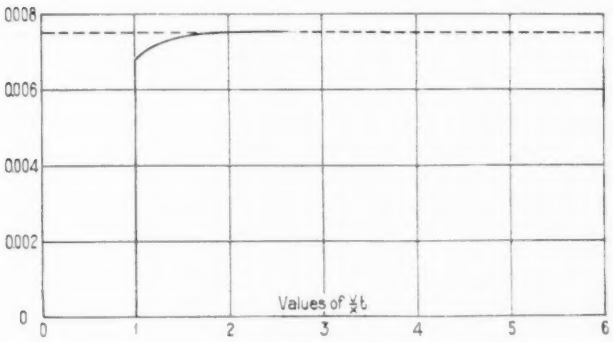


Fig. 18—Propagated voltage in line; $\frac{R}{2}\sqrt{\frac{C}{L}}\, x = ax = 3;\ \frac{G}{2}\sqrt{\frac{L}{C}}\, x = bx = 2.$

$J(t)$ and $W(t)$ respectively, while the first component of (210-b) and (211-b) introduce merely a delay. Thus, if the e.m.f. impressed at time $t=0$ is $f(t)$, the corresponding waves for $t \geq x/v$ or $x \geq x$, are

$$I = \sqrt{\frac{C}{L}} e^{-ax} f(t-x/v) + \int_{x/v}^{t} f(t-t_1) J'(t_1) dt_1, \tag{212}$$

$$V = e^{-ax} f(t-x/v) + \int_{x/v}^{t} f(t-t_1) W'(t_1) dt_1, \tag{213}$$

where $J'(t) = \frac{d}{dt} J(t)$ and $W'(t) = \frac{d}{dt} W(t)$.

The integrals of (212) and (213) can be computed and analyzed in precisely the same way as discussed in connection with the non-inductive cable problem, and are of very much the same character as the alternating current waves of the cable. In the total waves, however, as given by (212) and (213), a very essential difference is introduced by the absence of the first terms, which represent undistorted waves propagated with velocity v. Thus, if the impressed e.m.f. is $\sin \omega t$, (212) and (213) become

$$I = \sqrt{\frac{C}{L}} e^{-ax} \sin \omega(t-x/v) + \int_{x/v}^{t} \sin \omega(t-t_1) J'(t_1) dt_1, \text{ for } t \geq x/v \tag{214}$$

$$V = e^{-ax} \sin \omega(t-x/v) + \int_{x/v}^{t} \sin \omega(t-t_1) W'(t_1) dt_1, \text{ for } t \geq x/v. \tag{215}$$

Now the first terms of (214) and (215) are simply the usual steady-state expressions for the current and voltage waves when the frequency is sufficiently high to make the steady-state attenuation constant equal to a and the phase velocity equal to v. Furthermore the integral terms become smaller and smaller as the applied frequency $\omega/2\pi$ is increased. It follows, therefore, that for high frequencies the waves assume substantially their final steady value at time $t=x/v$, and that the tails of the waves, or the transient distortion, becomes negligible. This is a consequence entirely of the

presence of inductance in the line, and shows its extreme importance in the propagation of alternating waves and the reduction of transient distortion.

It should be pointed out, however, that if the line is very long and the attenuation is very high, the integral terms of (214) and (215) are not negligible unless the applied frequency is correspondingly very high. For example, on a long submarine cable, the a.c. attenuation is so large that the first terms of (214) and (215) are very small, and $J(t)$ is very large compared with $\sqrt{C/L}\ e^{-ax}$. Consequently here there is very serious transient distortion and alternating currents are therefore not adapted for submarine telegraph signalling.

This discussion may possibly be made a little clearer, without detailed analysis, if we recall the discussion of alternating current propagation in the non-inductive cable of the preceding chapter. From that analysis it follows that, when the applied frequency $\omega/2\pi$ is sufficiently high, the integral term of (214) becomes approximately

$$\frac{1}{\omega} J'(t)$$

and the complete current wave is

$$\sqrt{\frac{C}{L}} e^{-ax} \sin \omega(t-x/v) + \frac{1}{\omega} J'(t) \tag{216}$$

and similarly the voltage wave is

$$e^{-ax} \sin \omega(t-x/v) + \frac{1}{\omega} W'(t). \tag{217}$$

Now if the total attenuation ax is large the last terms of (216) and (217), before they ultimately die away, may become very large compared with the first terms, which represent the ultimate steady-state.

Appendix to Chapter VII. Derivation of Formula (211)

The only troublesome question involved in deriving (211) from (207) and (209) is that we have to differentiate with respect to x, in accordance with (207), the discontinuous function $F(t)$. To accomplish this we write (209) in the form

$$F(t) = \phi(t-x/v) e^{-\rho t} I_o(\sigma \sqrt{t^2 - x^2/v^2}) \tag{209-a}$$

where $\phi(t)$ is defined as a function which is zero for $t < x/v$ and unity for $t \geq x/v$. Clearly this is equivalent to (209) and permits us to deal

with $F(t)$ as a *continuous* function. Now, in accordance with (207), perform the operation of differentiation upon (209-a): we get

$$-v\frac{\partial F}{\partial x} = \frac{\partial}{\partial t}\phi(t-x/v)e^{-pt}I_o(\sigma\sqrt{t^2-x^2/v^2})$$

$$-v\phi(t-x/v)\frac{\partial}{\partial x}e^{-pt}I_o(\sigma\sqrt{t^2-x^2/v^2}).$$

The first expression follows from the fact that

$$\frac{\partial}{\partial x}\phi(t-x/v) = -\frac{1}{v}\frac{\partial}{\partial t}\phi(t-x/v).$$

We observe also that $\frac{\partial}{\partial t}\phi(t-x/v)=0$ except at $t=x/v$, when it is infinite. We also observe that, for $t \geq x/v$,

$$\int_0^t \frac{\partial}{\partial t}\phi(t-x/v)dt = 1$$

and that the whole contribution to the integral occurs at $t=x/v$. With these points clearly in mind, the expression

$$V = -v\int_0^t \frac{\partial F}{\partial x}dt$$

reduces to (211) without difficulty.

CHAPTER VIII

Propagation of Current and Voltage in Artificial Lines and Wave Filters

The artificial line here considered is a periodic structure, composed of a series of sections connected in tandem, each section consisting of a lumped impedance z_1 in series with the line, and a lumped

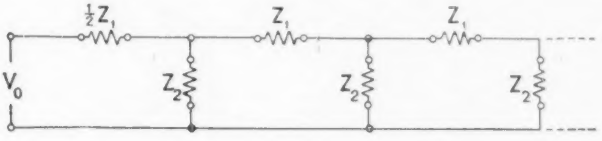


Fig. 19

impedance z_2 in shunt across the line. In the artificial line which we shall consider it will be assumed that the voltage is applied at the middle of the initial or zeroth section, as shown in Fig. 19. This termination is chosen because of its practical importance, and be-

cause also of the fact that the mathematical analysis is simplified thereby. Furthermore any other termination can be regarded and dealt with as an additional terminal impedance, so there is no essential loss of generality involved.

A study of the properties of the artificial line is of practical importance for several reasons:

1. The artificial line is often used as a model of an actual transmission line and it is therefore of importance to determine theoretically the degree of correspondence between the two.

2. The solution for the corresponding transmission line with continuously distributed constants is derivable from the solution for the artificial line by keeping the total inductance, resistance, capacity and leakage constant or finite, and letting the number of sections approach infinity.

3. The artificial line is very closely related, in its properties and performance, to the periodically loaded line, and its solution is, to a first approximation, a working solution for the loaded line.

4. The structure is of great importance in its own right, and when the impedance elements are properly chosen, constitutes a "wave filter."

We shall now derive the operational and symbolic equations which formulate the propagation phenomena in the artificial line. Let I_n denote the mesh current in the nth section of the line; I_{n-1} the mesh current in the $(n-1)^{\text{th}}$ section, etc. Now write down the expression for the voltage drop in the n^{th} section; in accordance with Kirchhoff's law we get:

$$(z_1+2z_2)I_n-z_2(I_{n-1}+I_{n+1})=0 \tag{218}$$

where, of course, the impedances have the usual significance.

Now this is a difference equation, as distinguished from a differential equation, but the method of solution is essentially the same. We *assume* a solution of the form

$$I_n=Ae^{-n\Gamma}+Be^{n\Gamma} \tag{219}$$

where A, B and Γ are independent of n, and substitute in (218). After some simple rearrangements we get

$$\{(z_1+2z_2)-2z_2\cosh\Gamma\}\{Ae^{-n\Gamma}+Be^{n\Gamma}\}=0. \tag{220}$$

Equation (218) is clearly satisfied by the assumed form of solution, and furthermore leaves the constants A and B arbitrary and at our

disposal to satisfy any boundary conditions, provided Γ is so chosen that

$$\begin{aligned}\cosh \Gamma &= \frac{z_1+2z_2}{2z_2} \\ &= 1+2\rho\end{aligned} \tag{221}$$

where $\rho = z_1/4z_2$.

Now by reference to equation (219) it is easily seen that Γ is the *propagation constant* of the artificial line, precisely analogous to the propagation constant γ of the smooth line. In terms of the impedances z_1 and z_2, the propagation constant of the artificial line is determined by (221). This equation may either be regarded as an operational equation or a symbolic equation, depending on whether the impedances are expressed in terms of the operator p or in terms of $i\omega$, where ω is 2π times the frequency.

Now suppose in (221) we write $e^{\Gamma} = x$; the equation becomes

$$x + 1/x = 2(1+2\rho)$$

and solving for x we get

$$\begin{aligned}x = e^{\Gamma} &= (1+2\rho) + \sqrt{(1+2\rho)^2 - 1} \\ &= (\sqrt{1+\rho} + \sqrt{\rho})^2 = (\sqrt{1+\rho} - \sqrt{\rho})^{-2}\end{aligned} \tag{222}$$

which is an explicit formula for Γ.

Now return to equation (219) and let us assume that the line is either infinitely long, or, what amounts to the same thing, that it is closed by an impedance which suppresses the reflected wave. We assume also that a voltage V_0 is impressed at mid-series position of the zeroth section ($n = $). Equation (219) becomes

$$I_n = Ae^{-n\Gamma}$$

and the currents in the zeroth and 1st sections are

$$I_0 = A, \qquad I_1 = Ae^{-\Gamma}.$$

Now, by direct application of Kirchhoff's law to the zeroth section, we have

$$V_0 = (\tfrac{1}{2}z_1 + z_2)I_0 - z_2 I_1,$$

whence

$$A\{\tfrac{1}{2}z_1 + z_2(1-e^{-\Gamma})\} = V_0. \tag{223}$$

But

$$I_0 = A = \frac{1}{K}V_0,$$

$$I_n = \frac{V_0}{K}e^{-n\Gamma},$$

where K is the *characteristic impedance* of the artificial line (at mid-series position). Hence by (223) and (222)

$$\frac{1}{K}=\frac{1}{z_2}\frac{1}{(1-e^{-\Gamma})+2\rho}$$
$$=\frac{1}{2z_2}\frac{1}{\sqrt{\rho+\rho^2}}=\frac{1}{\sqrt{z_1z_2}}\frac{1}{\sqrt{1+\rho}}. \tag{224}$$

By aid of the preceding the direct current wave can be written as

$$I_n=\frac{V_0}{\sqrt{z_1z_2}}\left[\frac{\sqrt{1+\rho}-\sqrt{\rho}}{\sqrt{1+\rho}}\right]^{2n}. \tag{225}$$

This formula is not so physically suggestive as its equivalent

$$I_n=\frac{V_0}{K}e^{-n\Gamma}$$

but is useful when we come to the solution of the operational equation.

Before proceeding with the operational equation, and the investigation of transient phenomena in artificial lines, it will be of interest to deduce from the foregoing the unique and remarkable properties of wave filters in the steady state. For this purpose we return to equation (221)

$$\cosh\Gamma=1+2\rho.$$

Now suppose that the series impedance z_1 is an inductance L and the shunt impedance z_2 a capacity C, so that, symbolically,

$$z_1=i\omega L,\ z_2=\frac{1}{i\omega C},\ \rho=-\frac{\omega^2LC}{4},$$

and

$$\cosh\Gamma=1-\tfrac{1}{2}\,\omega^2LC. \tag{226}$$

Now let us write $\Gamma=i\theta$, where $i=\sqrt{-1}$; the preceding equation becomes

$$\cos\theta=1-\tfrac{1}{2}\,\omega^2LC \tag{227}$$

and the *ratio of currents* in adjacent sections is $e^{-i\theta}$. *Consequently if θ is a real quantity the ratio of the absolute values of the currents in adjacent sections is unity, and the current is propagated without attenuation.*

Inspection of equation (227) shows that θ is real provided the right hand side lines between $+1$ and -1: or that ω lies between 0 and $2/\sqrt{LC}$. Consequently this type of artificial line transmits, in the steady state, sinusoidal currents of all frequencies from zero to $1/\pi\sqrt{LC}$ without attenuation. It is known as the low-pass filter.

If we invert the structure, that is, make the series impedance z_1 a capacity C and the shunt impedance z_2 an inductance L, so that

$$z_1=\frac{1}{i\omega C},\quad z_2=i\omega L,\quad \rho=-\frac{1}{4\omega^2LC},$$

we get, corresponding to (226) and (227),

$$\cosh\Gamma=1-\frac{1}{2\omega^2LC}, \tag{228}$$

$$\cos\theta=1-\frac{1}{2\omega^2LC}. \tag{228a}$$

This type of artificial line transmits without attenuation currents of all frequencies for which the right hand side of (228-a) lies between $+1$ and -1; that is, all frequencies from infinity to a lower limiting frequency $1/4\pi\sqrt{LC}$, while it attenuates all frequencies below this range. It is known, on this account, as the high-pass filter.

It is possible by using more complicated impedances to design filters which transmit a series of bands of frequencies. We cannot, however, go into the complicated theory of wave filters here, which has been covered in a series of important papers. One point should be noted, however: transmission without attenuation implies that the impedance elements are non-dissipative. Actually, of course, all the elements introduce some loss, so that in practice the filter attenuates all frequencies. Careful design, however, keeps the attenuation very low in the transmission bands.

We shall now derive the indicial admittance formulas for some representative types of artificial lines and wave filters from the operational formula

$$A_n=\frac{1}{\sqrt{(1+\rho)z_1z_2}}\left[\sqrt{1+\rho}+\sqrt{\rho}\right]^{-2n}. \tag{229}$$

This equation follows directly from (225) on putting $V_0=1$.

We start with the so-called low-pass filter on account of its simplicity and also its great importance in technical applications. This type of filter consists of series inductance L and shunt capacity C. The general case which includes series resistance R and shunt leakage G has been worked out (see Transient Oscillations, Trans. A. I. E. E., 1919). The solution is, however, extremely complicated and will not be dealt with here. We shall, instead, consider the important and illuminating case where the series and shunt losses are so related

as to make the circuit quasi-distortionless. We therefore take, operationally,

$$\begin{aligned} z_1 &= pL+R = L(p+\lambda) \\ 1/z_2 &= pC+G = C(p+\lambda) \end{aligned} \tag{230}$$

where $\lambda = R/L = G/C$.

We then have

$$\begin{aligned} z_1 z_2 &= L/C, \\ z_1/z_2 &= LC(p+\lambda)^2, \\ \rho &= \frac{LC}{4}(p+\lambda)^2. \end{aligned} \tag{231}$$

Now by reference to formula (229) we see that A_n is a function of $(p+\lambda)$; thus

$$A_n = \frac{1}{Z_n(p+\lambda)} = \left(1+\frac{\lambda}{\rho}\right)\frac{\rho}{(\rho+\lambda)Z_n(\rho+\lambda)}.$$

Now write

$$A_n^0 = \frac{1}{Z_n(p)}.$$

It follows at once from reference to theorem VII that

$$A_n = \left(1+\lambda\int_0^t dt\right)A_n^0 e^{-\lambda t} \tag{232}$$

so that the problem is reduced to the solution of the operational equation for A_n^0. Writing $\omega_c = 2/\sqrt{LC}$, we have

$$\begin{aligned} A_n^0 &= \sqrt{\frac{C}{L}}\frac{1}{\sqrt{1+(p/\omega_c)^2}}\left[\sqrt{1+(p/\omega_c)^2}+p/\omega_c\right]^{-2n} \\ &= \sqrt{\frac{C}{L}}\frac{\omega_c}{\sqrt{p^2+\omega_c^2}}\left[\frac{\sqrt{p^2+\omega_c^2}-p}{\omega_c}\right]^{2n}. \end{aligned} \tag{233}$$

Now refer to formula (n) of the table of integrals; writing $\sqrt{L/C}=k$, we see by Theorem V that

$$A_n^0 = \frac{1}{k}\int_0^{\omega_c t} J_{2n}(\tau)\,d\tau \tag{234}$$

where $J_{2n}(\tau)$ is the Bessel function of order $2n$ and argument τ. We note also that this is the indicial admittance of the non-dissipative low-pass wave filter; that is, the current in the n^{th} section in response

to a unit e.m.f. applied to the initial section ($n=0$). From (232) and (234) it follows at once that

$$A_n = e^{-\lambda t}\frac{1}{k}\int_0^{\omega_c t} J_{2n}(\tau)d\tau + \frac{\lambda}{k}\int_0^t d\tau e^{-\lambda\tau}\int_0^{\omega_c\tau} J_{2n}(\tau_1)d\tau_1.$$

Integrating the second member by parts and noting that $A_n(o)=0$, this reduces to

$$A_n = \frac{1}{k}\int_0^{\omega_c t} e^{-\frac{\lambda}{\omega_c}\tau} J_{2n}(\tau)d\tau \tag{235}$$

which is the indicial admittance formula for the quasi-distortionless low-pass filter, or artificial line.

Before discussing these formulas, it is of interest to derive the formula for A_n'' by power series expansion. Formula (233) can be written

$$A_n'' = \frac{1}{k}\left(\frac{\omega_c}{p}\right)^{2n+1}\frac{1}{\sqrt{1+(\omega_c/p)^2}\left[1+\sqrt{1+(\omega_c/p)^2}\right]^{2n}}.$$

This can be expanded in a series in inverse powers of p; thus

$$A_n'' = \frac{1}{k2^{2n}}\left[\left(\frac{\omega_c}{p}\right)^{2n+1} - \frac{2n+2}{2^2 1!}\left(\frac{\omega_c}{p}\right)^{2n+3} + \frac{(2n+3)(2n+4)}{2^2 2!}\left(\frac{\omega_c}{p}\right)^{2n+5} - \cdots\right].$$

Replacing $1/p^n$ by $t^n/n!$ in accordance with the Heaviside Rule we get

$$A_n'' = \frac{2}{k}\left[\frac{1}{(2n+1)!}\left(\frac{\omega_c t}{2}\right)^{2n+1} - \frac{2n+2}{1!(2n+3)!}\left(\frac{\omega_c t}{2}\right)^{2n+3} + \frac{(2n+3)(2n+4)}{2!(2n+5)!}\left(\frac{\omega_c t}{2}\right)^{2n+5} - \cdots\right]. \tag{235-a}$$

This can be recognized as the power series expansion of (234).

The *artificial cable* is also of interest and practical importance. In this structure the series impedance is a resistance R and the shunt impedance is a capacity C, so that

$$\begin{aligned} &z_1 = R,\ 1/z_2 = pC, \\ &z_1 z_2 = R/pC,\quad z_1/z_2 = pRC, \\ &\rho = pRC/4. \end{aligned} \tag{236}$$

Now let us return to formula (229), and expand in inverse powers of ρ: we get

$$A_n = \frac{1}{2^{2n}\sqrt{\rho z_1 z_2}^n}\left\{\frac{1}{\rho^n} - \frac{2n+2}{2^2 1!}\frac{1}{\rho^{n+1}} + \frac{(2n+3)(2n+4)}{2^2 2!}\frac{1}{\rho^{n+2}} - \cdots\right\} \quad (237)$$

Now since $\sqrt{\rho z_1 z_2} = \frac{R}{2}$, we have

$$A_n = \frac{2}{2^n R}\left\{\left(\frac{2}{RCp}\right)^n - \frac{2n+2}{2.1!}\left(\frac{2}{RCp}\right)^{n+1} + \frac{(2n+3)(2n+4)}{2^2 2!}\left(\frac{2}{RCp}\right)^{n+2} - \cdots\right\}.$$

Replacing $1/p^n$ by $t^n/n!$ we get finally

$$A_n = \frac{2}{2^n R}\left\{\frac{1}{n!}\left(\frac{2t}{RC}\right)^n - \frac{(2n+2)}{2.1!(n+1)!}\left(\frac{2t}{RC}\right)^{n+1} + \frac{(2n+3)(2n+4)}{2^2.2!(n+2)!}\left(\frac{2t}{RC}\right)^{n+2} - \cdots\right\}. \quad (238)$$

For large values of n and t this series is difficult to compute or interpret. It can, however, be recognized as the series expansion of the function

$$A_n = \frac{2}{R}e^{-\frac{2t}{RC}} I_n\left(\frac{2t}{RC}\right) \quad (239)$$

where $I_n(2t/RC)$ is the Bessel function I_n of order n and argument $(2t/RC)$. This solution, it may be remarked, can be derived directly by a modification of the integral formula (n).

It is beyond the scope of this paper to consider other types of artificial lines and wave filters; for a fairly extensive discussion the reader is referred to "Transient Oscillations in Electric Wave-Filters," B. S. T. J., July, 1923. The low-pass wave filter, however, both in its own right and on account of its close relation to the periodically loaded line, deserves further discussion.

For the non-dissipative low-pass wave filter, we have

$$A_n^0 = \frac{1}{k}\int_0^{\omega_c t} J_{2n}(\tau)\,d\tau \quad (234)$$

while for the quasi-distortionless low-pass wave filter

$$A_n = \frac{1}{k}\int_0^{\omega_c t} e^{-\mu\tau} J_{2n}(\tau)\,d\tau \quad (235)$$

where $\mu = \lambda/\omega_c = R/L\omega_c = R/2vL$.

Computation and analysis of these formulas involve an elementary knowledge of Bessel functions. The properties necessary for our purposes are briefly discussed in an appendix to this chapter.

The indicial admittances for the non-dissipative low-pass filter,

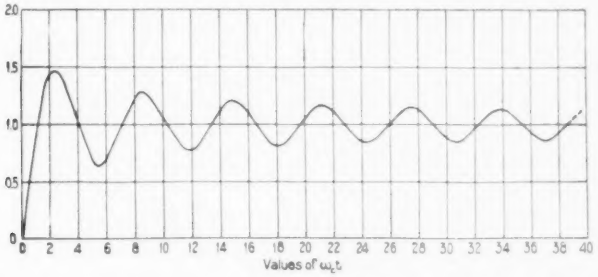


Fig. 20—Low pass wave filter. Indicial admittance of initial section ($n=0$). Multiply ordinates by $\sqrt{C/L}$

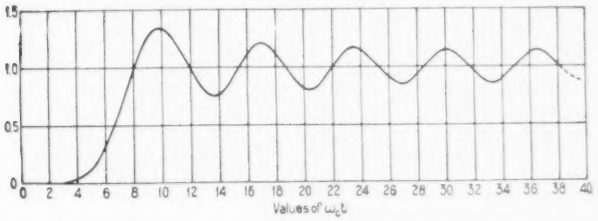


Fig. 21—Low pass wave filter. Indicial admittance of third section ($n=2$). Multiply ordinates by $\sqrt{C/L}$

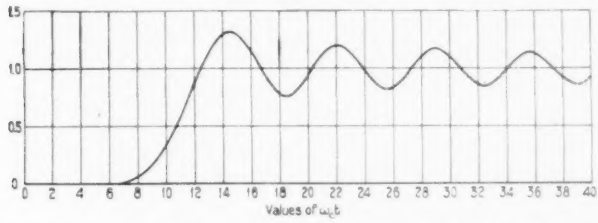


Fig. 22—Low pass wave filter. Indicial admittance of fifth section ($n=4$). Multiply ordinates by $\sqrt{C/L}$

that is, the current in response to a steady unit e.m.f. applied at time $t=0$, are shown in the curves of Figs. 20, 21 and 22, for the initial or zeroth, the 3rd and the 5th sections, respectively. These curves together with the exact and approximate formulas given

above are sufficient to give a reasonably comprehensive idea of the general character of these oscillations and their dependence on the number of sections and the constants of the filter.

It will be observed that the current is small until a time approximately equal to $2n/\omega_c = n\sqrt{L_1C_2}$ has elapsed after the voltage is applied. Consequently the low-pass filter behaves as though currents were transmitted with a finite velocity of propagation $\omega_c/2 = 1/\sqrt{L_1C_2}$ sections per second. This velocity is, however, only apparent or virtual since in every section the currents are actually finite for all values of time >0.

After time $t = n\sqrt{L_1C_2}$ has elapsed the current oscillates about the value $1/k$ with increasing frequency and diminishing amplitude. The amplitude of these oscillations is approximately

$$\frac{1/k}{\sqrt{1-(2n/\omega_c t)^2}}\sqrt{\frac{2}{\pi\omega_c t}}$$

and their instantaneous frequency (measured by intervals between zeros)

$$\frac{\omega_c}{2\pi}\sqrt{1-(2n/\omega_c t)^2}.$$

The oscillations are therefore ultimately of cut-off or critical frequency $\omega_c/2\pi$ in all sections, but this frequency is approached more and more slowly as the number of filter sections is increased.

Figs. 23, 24, 25, give the indicial admittance in the 100th, 500th and 1000th section of the low-pass wave filter. The filter itself seldom

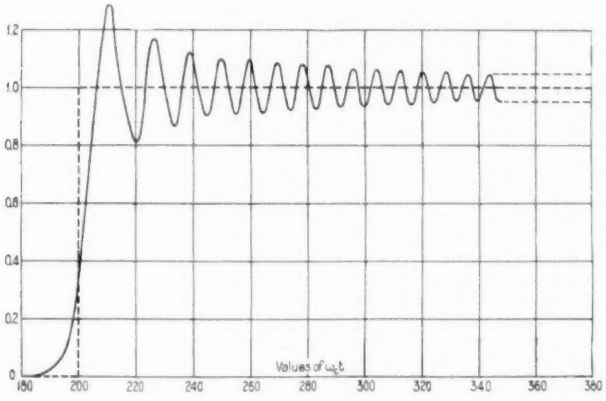


Fig. 23—Low pass wave filter. Indicial admittance of 100th section ($n=99$). Multiply ordinates by $\sqrt{C/L}$

embodies more than 5 sections. The case of a large number of sections is of interest, however, because it represents a first approximation to the periodically loaded line. While the non-dissipative

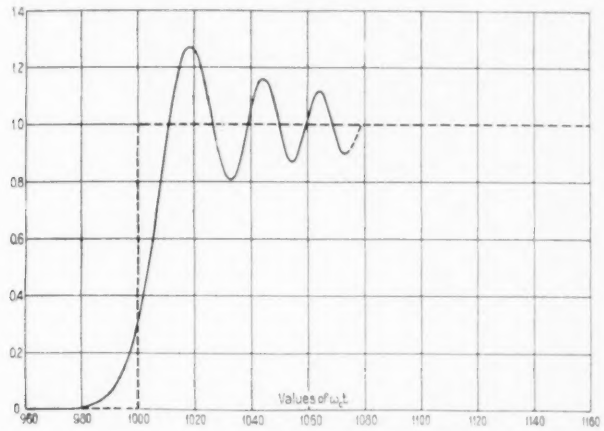


Fig. 24—Low pass wave filter. Indicial admittance of 500th section ($n=499$). Multiply ordinates by $\sqrt{C/L}$.

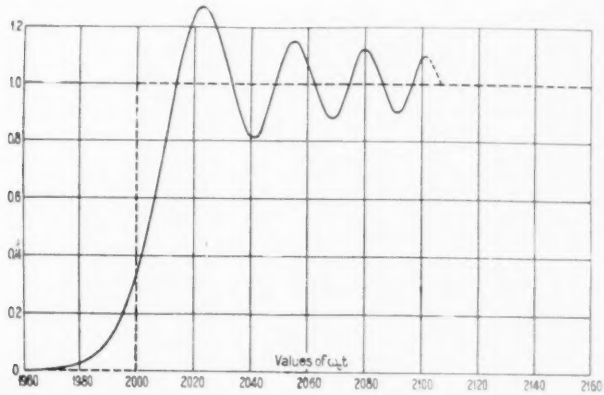


Fig. 25—Low pass wave filter. Indicial admittance of 1000th section ($n=999$). Multiply ordinates by $\sqrt{C/L}$.

line is ideal and unrealizable, its study is of practical importance because in this type of line the effect of the discontinuous character of the loading of the periodically loaded line is isolated and exhibited in the clearest possible manner.

The dotted curves represent the current in the corresponding smooth line. For the smooth line, the current, as we have seen, is discontinuous, being identically zero for a time $vt = n$ and having an instantaneous jump to its final value $\sqrt{C/L}$ at $vt = n$. The current in the artificial or periodically loaded line differs from that in the corresponding smooth line in three important respects: (1) the absence of the abrupt discontinuous wave front, (2) the presence of superposed oscillations, and (3) *the absence of a true finite velocity of propagation*. It will be observed, however, that the current in any section is negligibly small or even sensibly zero until $vt = n$, so that the current is propagated with a *virtual* velocity $1/\sqrt{LC}$ per section. The presence of a well marked wave front is also evident although this is not abrupt, as in the smooth line. The effective slope of the wave front becomes smaller as the current wave travels out on the line, decreasing noticeably as the number of sections is increased. When the number of sections becomes large, however, the decrease in the slope is not rapid, being in the 500th section about 60 per cent. of that in the 100th section.

The superposed oscillations are of interest. These are initially of a frequency depending upon and decreasing with the number of sections, n, but in all sections ultimately attaining the frequency

$$\frac{1}{\pi\sqrt{LC}} = \frac{v}{\pi}$$

which is the critical or cut-off frequency of the line, above which steady-state currents are attenuated during transmission and below which they are unattenuated. When vt is large compared with n the amplitude of these oscillations becomes $\sqrt{1/\pi vt}$ so that they ultimately die away and the current approaches the value $\sqrt{C/L}$ for all sections. The current in the loaded line is thus asymptotic to the current in the corresponding smooth line and oscillates about it with diminishing amplitude and increasing frequency.

Since the abscissas of these curves represent values of $2vt = 2t/\sqrt{LC}$, and the ordinates are to be multiplied by $\sqrt{C/L}$ to translate into actual values, the curves are of universal application for all values of the constants L and C.

The investigation of the building-up of alternating currents in wave filters and loaded lines is very important. It depends for the non-dissipative case on the properties of the definite integrals

$$\int_0^{\omega_c t} \sin w\tau \, J_n(\tau) d\tau,$$

$$\int_0^{\omega_c t} \cos w\tau \, J_n(\tau) d\tau,$$

where $w=\omega/\omega_c$ and $\omega=2\pi$ times the applied frequency. The mathematical discussion is, however, quite complicated and will not be entered into here. The reader, who wishes to follow this further, is referred to Transient Oscillations, Trans. A. I. E. E., 1919 and Transient Oscillations in Electric Wave Filters, B. S. T. J., July, 1923.

Appendix to Chapter VIII. Note on Besel Functions

The Bessel Functions of the first kind, $J_n(x)$ and $I_n(x)$, are defined, when n is zero or a positive integer, by the absolutely convergent series

$$J_n(x)=\frac{x^n}{2^n.n!}\left\{1-\frac{x^2}{2(2n+2)}+\frac{x^4}{2.4(2n+2)(2n+4)}-\frac{x^6}{2.4.6.(2n+2)(2n+4)(2n+6)}+\cdots\right\},$$

$$I_n(x)=\frac{x^n}{2^n.n!}\left\{1+\frac{x^2}{2(2n+2)}+\frac{x^4}{2.4(2n+2)(2n+4)}+\frac{x^6}{2.4.6.(2n+2)(2n+4)(2n+6)}+\cdots\right\}.$$

In the following discussion of the properties of these functions it will be assumed that the argument x is a pure real quantity.

For large values of the argument (x large compared with n), the behavior of the functions is shown by the asymptotic expansions: —

$$I_n(x)=\frac{e^x}{\sqrt{2\pi x}}\left\{1-\frac{4n^2-1}{1!\,(8x)}+\frac{(4n^2-1)(4n^2-9)}{2!\,(8x)^2}-\frac{(4n^2-1)(4n^2-9)(4n^2-25)}{3!\,(8x)^3}+\cdots\right\},$$

$$J_n(x)=\sqrt{\frac{2}{\pi x}}\left\{P_n\cos\left(x-\frac{2n+1}{4}\pi\right)-Q_n\sin\left(x-\frac{2n+1}{4}\pi\right)\right\},$$

where

$$P_n=1-\frac{(4n^2-1)(4n^2-9)}{2!\,(8x)^2}+\frac{(4n^2-1)(4n^2-9)(4n^2-25)(4n^2-49)}{4!\,(8x)^4}-\cdots,$$

$$Q_n=\frac{4n^2-1}{8x}-\frac{(4n^2-1)(4n^2-9)(4n^2-25)}{3!(8x)^3}+\cdots.$$

We thus see that I_n increases indefinitely and behaves ultimately as

$$\frac{e^x}{\sqrt{2\pi x}}.$$

The function $J_n(x)$, however, is oscillatory and ultimately behaves as

$$\sqrt{\frac{2}{\pi x}}\cos\left(x-\frac{2n+1}{4}\pi\right).$$

For all orders of n

$$\int_0^\infty J_n(x)dx=1.$$

The properties of $J_n(x)$ may be described qualitatively as follows:—

When the argument is less than the order ($0\leqq x<n$) the function is very small and positive, and is initially zero (except when $n=0$). In the neighborhood of $x=n$, the function begins to build up and reaches a maximum a little beyond the point $x=n$. Thereafter the function oscillates with increasing frequency and diminishing amplitude, and ultimately behaves as

$$\sqrt{\frac{2}{\pi x}}\cos\left(x-\frac{2n+1}{4}\pi\right).$$

When $n=0$, the initial value is unity, but the subsequent behavior of the function is as described above.

A more precise description of the function is gotten from the following approximate formulas.

$$J_n(x)\doteq B_n(x)\cos\Omega_n(x), \qquad \text{for } x>n$$

where

$$B_n(x)=\sqrt{\frac{2}{\pi x}}\left(1-\frac{m^2}{x^2}+\frac{3}{2}\frac{m^2}{x^4}\frac{1}{(1-m^2/x^2)^2}\right)^{1/4},$$

$$\Omega_n(x)=x\left[\sqrt{1-\frac{m^2}{x^2}}+\frac{m}{x}\sin^{-1}\left(\frac{m}{x}\right)-\frac{m^2}{4x^4}\frac{1}{(1-m^2/x^2)^{3/2}}\right]-\frac{2n+1}{4}\pi,$$

$$\Omega_n'(x)=\frac{d}{dx}\Omega_n(x),$$

$$=\sqrt{1-\frac{m^2}{x^2}+\frac{3}{2}\frac{m^2}{x^4}\frac{1}{(1-m^2/x^2)^2}},$$

and

$$m^2=n^2-1/4.$$

This approximate formula is valid only where $x>n$, its accuracy increasing with x and with n. For all orders of n it is quite accurate beyond the first zero of the function.

The "instantaneous frequency" of oscillation is approximately

$$\frac{1}{2\pi}\Omega_n'(x) = \frac{1}{2\pi}\sqrt{1-\frac{m^2}{x^2}+\frac{3}{2}\frac{m^2}{x^4}\frac{1}{(1-m^2/x^2)^2}}$$

By this it is meant that at any point $x(x>n)$ the interval between successive zeros is approximately $\pi/\Omega'(x)$. Otherwise stated, in the neighborhood of any point x, the function behaves like a sinusoid of amplitude $B_n(x)$ and frequency $\omega/2\pi$ where $\omega=\Omega_n'(x)$.

The following approximate formulas, while not sufficiently precise for the purposes of accurate computation except for quite large values of x, clearly exhibit the character of the functions for values of the argument $x>n$, and of the order $n>2$.

$$J_n(x) \doteq h_n\sqrt{\frac{2}{\pi x}}\cos\ (q_n x-\theta_n),$$

$$J_n'(x) = -q_n h_n\sqrt{\frac{2}{\pi x}}\sin\ (q_n x-\theta_n),$$

$$\int_0^x J_n(x)dx = 1+\frac{h_n}{q}\sqrt{\frac{2}{\pi x}}\sin\ (q_n x-\theta_n),$$

where

$$h_n=\left(\frac{1}{1-n^2/x^2}\right)^{1/4}=1+\frac{n^2}{4x^2},$$

$$q_n=\sqrt{1-n^2/x^2},$$

$$\theta_n=\frac{2n+1}{4}\pi-n\sin^{-1}(n/x),$$

Some Contemporary Advances in Physics—X
The Atom-Model, Third Part

By KARL K. DARROW

M. A Very Brief Recapitulation of What Has Gone Before

ABUNDANT evidence of many kinds exists to show that each and every distinct sort of atom is especially adapted to possess energy, not in any random quantity whatsoever, but in certain peculiar, definite, characteristic amounts. An atom having energy in one of these particular amounts apparently cannot add arbitrary quantities to its store, nor yield up arbitrary quantities from it; whenever the atom receives or whenever it gives energy, it receives or gives only just so much as is exactly sufficient to raise or reduce its supply to some one among the others of these distinctive quotas. For each of the chemical elements there is a great system of these distinctive energy-values. They are determined chiefly by analyzing spectra, and for most of the elements—the exceptions being those of which the spectra are excessively complicated—many of them have been evaluated very accurately and set down in tables. The system of distinctive energy-values for any element is a very important feature of that element; perhaps, indeed, the most important feature of all.

It is customary to say that when an atom acquires or surrenders energy, it passes from one into another state; the various states corresponding to its various distinctive energy-values are called its "Stationary States." This is a name which suggests, and is doubtless meant to suggest, that the energy-value of the atom is but one among many of its features, all of which change when the energy-value changes. This is a legitimate idea; theorizing about the atom consists in speculating about just such features. But the reader will go far and grievously astray if he lets the name signify to him that many of them are directly and definitely known. In some few cases there is good reason to believe that we know the magnetic moment of an atom in its normal state. Beyond these the energy-values are all that are known. If the reader chooses everywhere to replace "Stationary State" by "energy-value" he will be holding fast to the physical reality, to the one thing not liable to be compromised by the future trends of thought.

An atom may pass from one Stationary State to another because of colliding with an electron or another atom of the same or a different

kind; or by absorbing radiation it may pass from one Stationary State to another of higher energy; or it may pass spontaneously from one Stationary State to another of lower energy. In this last case, it emits radiation of which the frequency ν is related to the difference ΔU between the energy-values of the initial Stationary State and the final one by the equation

$$\nu = \Delta U/h, \qquad (h = 6.56.10^{-27} \text{ erg/sec.})$$

The same equation governs the last case but one, in which it connects the frequency of the absorbed radiation with the energy-difference between the two Stationary States from and into which the atom passes. On this equation is founded the method of analyzing spectra which is the most accurate and most widely applicable method of determining the energy-values of Stationary States. The other ways in which atoms are caused to pass from one State to another lead to methods of determining these states, which are almost useless for accurate measurements, but invaluable as controls.

The energy-values of the various Stationary States of an atom are interrelated, and sometimes it is possible to express a long sequence of them by means of a simple or a not very complicated formula. There are also interrelations between the distinctive energy-values for different elements; and this statement is meant to apply also to atoms from which electrons, one or more, have been detached, which should be considered as distinct though not as stable elements. There are unmistakable numerical relations among the Stationary States which come into being when atoms are subjected to electric or to magnetic fields. Finally there is the important principle that the spontaneous transitions between various pairs of states, which result in spectrum lines, do not occur equally often; and yet the relative oftenness or seldomness of their occurrence is itself regulated by laws. One finds instances in which transitions from a state A to a state B_1 are just twice as common as transitions from A to a state B_2 close to B_1. One finds instances in which transitions from a state A to a state B do not occur at all under usual conditions and an atom in state A cannot get into state B without touching at some state C from which A and B are both accessible. It is possible so to arrange the Stationary States of an atom that by looking at the situations of any two states in the arrangement, one can tell immediately whether direct transitions between them do occur or do not; and this arrangement is found to be suited to, even to be demanded by the numerical interrelations to which I alluded above. Upon these facts the classi-

fications of the Stationary States are founded, and the notations by which they are named.

The atom-model to which this article is devoted, the atom-model of Rutherford and Bohr, is designed to interpret these facts of the Stationary States, but not these alone. It is designed also to interpret certain experiments—chiefly, though not altogether, experiments on the deflections suffered by minute flying charged particles when they pass through matter—which indicate that an atom consists of a positively-charged nucleus with a congeries of electrons around it. Specifically, the results of these experiments agree with the notion that the Nth element of the Periodic Table consists of a nucleus with positive charge Ne and N electrons surrounding it; and this is the simplest and most satisfying notion with which they do agree. Yet there is something paradoxical about this atom-model; for electrons could neither stand still nor yet revolve permanently in orbits around a nucleus, if they conformed to the laws of electrostatics. Also there must needs be something paradoxical about any attempt to interpret the Stationary States by this model, for there is nothing inherent in it to make any energy-value preferable to any other. Under these circumstances Bohr's procedure was, resolutely to accept both paradoxes at once, and to say that the electrons can revolve permanently in those and just those particular orbits, whereby the energy of the atom assumes the particular values which are those of the observed Stationary States. This is easy to say; but it is not important, unless one succeeds in showing that those and only those particular orbits are set apart from all others by some peculiar feature, are distinguished by conforming to some particular principle, which can be exalted into a "Law of Nature" to complement or supersede the laws of electrostatics. Otherwise the atom-model would be of no value.

Thus in order to make the test of the atom-model, it is necessary to trace these orbits. One is confronted with this problem of orbit-tracing: *Given:* the observed energy-values of the Stationary States; *required:* to trace the orbits such that, when the electrons travel in them, the energy of the atom has these observed values. If this problem cannot be solved, it is impossible to take the next and essential step of ascertaining whether these particular orbits are distinguished in any particular way from all the other conceivable ones.

In the case of a single electron revolving about a nucleus, this problem is sometimes soluble. If the mass of the electron is regarded as invariable, and no outside influences are supposed to act upon the atom, then the solution is comparatively easy to attain. It was performed in the Second Part of this article. If an external magnetic

field is superposed, the problem is scarcely more difficult; if an external electric field is superposed, it is difficult but soluble—provided always that the mass of the electron be supposed invariable. If the mass of the electron is supposed to vary with its speed as the theory of relativity requires, and as certain experiments suggest, the problem remains soluble—provided that no outside influences act. For all these cases the orbits which yield the observed energy-values have been traced; and certain features have been shown to be common to all of these "permitted" orbits, and to no others, so that by these features the permitted orbits are set apart from all the rest. Inversely, anyone who is told what these features are, and who is sufficiently adept in dynamics, can trace all the orbits which display them and calculate the energy-values for these orbits and so predict the energy-values of all of the Stationary States of an atom consisting of a nucleus and a single electron. Such orbits are known as *quantized orbits*. The rules whereby they are set apart from all the multitude of orbits not permitted are the *Quantum Conditions*; which some one, it is to be hoped, will some day succeed in deriving from a general Principle of Quantization.

The most general way of phrasing these conditions is difficult to grasp, and the more intelligible ways are not the most general. The most general conditions yet formulated are not adequate for all cases; the completely adequate principle is yet to be discovered. For the purposes of this summary and of most of what follows, a very limited expression of the Quantum Conditions will be sufficient. In the Second Part of this article it was proved that the permitted orbits of an electron of invariable mass revolving in an inverse-square field such as is supposed to surround a bare nucleus, are certain ellipses. It was further stated, without proof, that if the electron of invariable mass revolves in a central field which deviates slightly from an inverse-square field, then the permitted orbits are certain "rosettes" or precessing ellipses—each orbit may be traced by imagining an ellipse revolving steadily in its own plane around the source of the central field (I will say the nucleus) at one of its foci. All the orbits are rosettes; the permitted orbits are certain rosettes which are distinguished from the others by a distinctive feature. One way of expressing this feature involves the angular momentum p_ϕ and the radial momentum p_r of the electron. In terms of the mass m of the electron, its distance r from the nucleus, and its angular velocity $d\phi/dt$ about the nucleus, these quantities are by definition:

$$p_r = m(dr/dt), \qquad p_\phi = mr^2(d\phi/dt). \tag{1}$$

The "principle of quantization" is, that the permitted orbits are marked out from all others in that they fulfil these conditions, which are the Quantum Conditions:

$$\int p_r \, dr = kh, \tag{2}$$

$$\int p_r \, dr + \int p_\phi \, d\phi = nh. \tag{3}$$

In these equations each integral is taken around one complete cycle of the corresponding variable; h stands for Planck's constant, and n and k take the values of all positive integers, k never surpassing n.

There is an alternative way of phrasing these quantum-conditions, which is much easier to visualize; but it emphasizes what are probably accidental features of the permitted rosettes, rather than fundamental ones. The rosettes are, as I have said, precessing ellipses; the major axes $2a$ and the minor axes $2b$ of these ellipses are, for the permitted rosettes

$$2a = n^2 h^2 / 2\pi^2 e^2 m, \tag{4}$$

$$2b = (k/n)\, 2a = k\, n\, h^2 / 2\pi^2 e^2 m, \tag{5}$$

in which n and k take as before the values of all positive integers, k not surpassing n.

Exactly the same principle governs the permitted orbits of an electron revolving in a perfect inverse-square central field, but varying in mass when its speed varies, as the theory of relativity requires. In this case also the orbits are rosettes, and the permitted orbits are particular rosettes set apart from all the others in that they fulfil (2) and (3), therefore automatically (4) and (5). The energy-values of these permitted orbits agree closely with those of the observed Stationary States of hydrogen and of ionized helium, the atoms of which are the only atoms believed to consist of a nucleus and one electron. Inversely, the orbits required to interpret the observed Stationary States are set apart from all the other conceivable orbits by the features expressed by (2) and (3), and by (4) and (5). On these close numerical agreements for hydrogen and ionized helium, and on other numerical agreements for the same atoms arising when external fields are applied, the prestige of Bohr's atom-model is founded.

The integers n and k, the *total quantum number* and the *azimuthal quantum number*, are used as indices to symbolize the various Stationary States of hydrogen and ionized helium to which they correspond. Thus the symbol "3_2" stands for the Stationary State of either atom,

which in the atom-model is realized when the electron circulates in a rosette, or precessing ellipse, for which $n=3$ and $k=2$. Orbits for which $k=n$ are circles; orbits for which $k<n$ are (precessing) ellipses, and the farther k falls below n the more eccentric (narrower) is the ellipse, although its major axis is independent of k. I have repeated all of these statements about precessing ellipses and their quantum-numbers, because a great part of the speculation about atoms possessed of more than one electron consists of persevering and obstinate attempts to interpret their behavior by as nearly as possible the same ideas.

It is essential to remember also that all the energy-values of the Stationary States are reckoned from the "State of the Ionized Atom," in which State the energy—i.e., the energy of a system composed of one atom deprived of an electron, and one electron far away—is equated to zero.

N. Introduction to the Speculations About Atoms with More than One Electron

All atoms, except those of hydrogen and ionized helium, possess more than one electron. There is much evidence of various kinds for this assertion; and certainly the spectra of these other atoms cannot be interpreted as those of the first two have been. Thus we are confronted with the problem of a system composed of a nucleus and more than one electron. The similarities between the spectra of hydrogen and ionized helium, and those of other elements, are important enough to make it desirable to use the same sort of explanation. We imagine the various electrons, when there are two or more, each to describe certain permitted orbits, set apart from the multitude of other conceivable orbits by peculiar features expressible by a Principle of Quantization.

Here at the outset we meet with the great hindrance to success in this problem. It is not possible to determine what features are common to permitted orbits, for it is not possible even to trace the permitted orbits. The general problem of tracing the paths of three or more bodies, attracting or repelling one another according to the inverse-square law, remains unsolved. Considering that for centuries the related but simpler problem of celestial mechanics has been under continual and powerful attack, the general problem may fairly confidently be regarded as insoluble. There is very little hope of ever dominating it to such an extent, that the spectra of atoms with two or more electrons can be interpreted exactly by Bohr's atom-model,

or can be used as strong support to that theory. If those two spectra of hydrogen and ionized helium were unknown, it is unlikely that the atom-model would ever have been suggested; it is more than unlikely that the atom-model could ever have been regarded as satisfactory. To this day the prestige of the atom-model results almost entirely from its achievement with those two spectra.

Why then trouble with applying it to the interpretation of other spectra? Several good reasons can be given. For instance, it may be that a system of several electrons about a nucleus acts in some respects as a unit—that its motion can be considered in some ways as the motion of a rigid body, that principles of quantization can be found for the system as a whole, similar to the principles used for quantizing the smaller and yet perhaps not more consolidated system which a single electron is. Here and there in the discussion we shall find indications that this way of thinking is suitable.

Again, one is justified in arguing that if in simple cases a certain law is proved, and if in complex cases neither that nor any other law can be proved nor disproved, than we should assume that the law proved for the simple cases extends over the complex ones. Few events in this world take place under such conditions that conservation of energy can be proved to prevail during them; yet, from the fact that conservation of energy has been verified in whatever events it has been tested, we do not hesitate to infer that it prevails in all. Bohr's model having been so strongly fortified by the data for the only two atoms for which it can be completely tested, why not assume it for the others?

And finally, there is the point that many of the data of experiment are almost universally expressed in terms of the model, so that the physical literature of today is almost incomprehensible without some knowledge of it. Unfortunate as this is, it shows that the model is a valuable aid for visualizing the facts. This justifies any model; but must not be construed as evidence for it.

It will be expedient to divide the subject substantially under these following headings.

(a) *The Helium Atom.* This, as the case of an atom composed presumably of a nucleus and two electrons, comes nearest to being amenable to calculation. Certain mechanically possible orbits of the two electrons, possessing the peculiar features of the "permitted" orbits of a single electron revolving around a nucleus, have been traced and their energy-values calculated. Not one of them has given the observed energy-value of a Stationary State of the helium atom. It is the consensus of opinion that whatever the features

which distinguish the permitted orbits may be, they are not those which prevail in the hydrogen atom.

(b) *Alkali-Metal Atoms.* For these there is reason to believe that one electron is normally located far beyond all the others, and may be supposed to revolve around a "residue" consisting of all the others and the nucleus. At a great distance, the field due to this residue will be very nearly a central field such as would surround a nucleus of charge $+e$,—a hydrogen nucleus; for, from a great distance, the nucleus and the electrons of the residue will seem almost to coincide in place. Nearer in, the forces due to the electrons of the residue may be supposed to compound with that due to the nucleus in such a way that a central field, not varying as the inverse square, results. Thus rosette orbits may be expected (for this reason I quoted the principle of quantization for such orbits in Section M). An enormous amount of effort has been spent in constructing central fields, such that the rosette orbits obeying the quantum-conditions (2) and (3) have nearly the energy-values which the Stationary States of these atoms are known to possess. Always, the emission of a spectrum line is supposed to result from a transition of the outermost or valence electron from one orbit to another, the electrons of the residue being scarcely or not at all affected. Such is the general explanation for the far-reaching and yet imperfect resemblance of the spectra of these metals to that of hydrogen.

(c) *Other Elements.* As one passes across the Periodic Table from left to right along any row, the spectra rapidly lose resemblance to the hydrogen spectrum. This is taken to mean that the assumption used for alkali metals—the assumption that one electron lies far beyond the others, and executes transitions while the others remain unaffected—departs progressively further from the truth. Evidence exists that simultaneous transitions of two electrons occur, and very likely yet more drastic rearrangements taking place *en bloc*.

(d) *Building of Atoms by Consecutive "Binding" of Electrons.* An atom composed, when complete, of Z electrons arranged about a nucleus bearing the charge $+Ze$, may have been formed originally in Z stages by the consecutive advent of Z electrons, the first annexing itself to the bare nucleus, the second joining itself to the system composed of the nucleus and the first, and so on until as many have arrived as the nucleus is able to hold. Each of these stages should be accompanied by the emission of lines belonging to a particular spectrum; the ordinary hydrogen spectrum accompanies the formation of a hydrogen atom by the step-by-step binding of an electron to a nucleus of charge e, the ionized-helium spectrum accompanies

the joining of the first electron to a nucleus of charge $+2e$, the neutral-helium spectrum the adhesion of the second electron to such a nucleus. Spectra corresponding to the latest four or five stages, in the formation of atoms having many electrons when completed, have been observed. To a certain extent, but not entirely, an atom with Z electrons and a nuclear charge Z resembles an atom with Z electrons and a nuclear charge $Z+1$. To a certain extent, therefore, each atom in the Periodic System may be regarded as resembling the last stage but one in the formation of the next following atom. This fact is important in the interpretations of the Periodic Table.

(e) *Multiplets.* We next take account of the fact that the sequences of Stationary States, mentioned in the elementary theory and description of spectra, are actually sequences of groups of Stationary States; and inquire what may be supposed to differentiate the several states of a group from one another. An elaborate formal theory is based on the assumption that all of the electrons of what I have called the "residue" of the atom revolve, if not literally as a rigid block, at least with a resultant angular momentum which itself is quantized; and that the outermost electron revolves in its own orbit around this residue, the different Stationary States of the group differing from one another in respect of the inclination of the orbit to the axis of rotation of the residue. The theory is not quite coherent with what has gone before; and for that reason the reader should try to separate its essential qualities from its accidental ones.

(f) *Magnetic Properties of Atoms.* A magnetic field should treat a system of electrons revolving around a nucleus in the same way as it treats one electron, as was said in the Second Part of this article. One would expect that in this case, if in any, the behavior of complex atoms would resemble that of the hydrogen atom; yet there is a striking and inexplicable contrast. This, like the spectrum of the helium atom, shows that either the quantum conditions governing the hydrogen atom are not universal, or the expressions hitherto found for the quantum conditions are too limited. From the responses of atoms to magnetic fields something is learned about the magnetic properties of atoms and their residues, some part of which can be tested by direct experiment; and these experiments include what are probably, all things considered, the most perplexing and fascinating ones of recent years.

(g) *Interpretation of X-ray Spectra.* X-ray spectra are analyzed as other spectra are, and each absorption and each emission of an X-ray by an atom is associated with a transition between two Stationary States; these "X-ray Stationary States" however are distin-

guished from the others, by the circumstance that every one of them involves the absence of an electron from the atom; consequently they may be described as Stationary States of an atom-residue. There is reason to believe that each distinct State involves the absence of a particular one, or of one out of a particular group, of the electrons bound to the nucleus during the earlier stages of the imagined building of the atom by successive "binding" of electrons. The speculations about X-ray spectra consist largely in attempts to correlate the individual States with absences of particular electrons.

O. The Helium Atom

The problem of the nucleus with two electrons, the "dilemma of the helium atom" as van Vleck calls it, is one of the most tantalizing in contemporary physics. One feels confident *a priori* that the same quantum conditions as suffice so beautifully to constrain the one-electron atom to yield the hydrogen spectrum should also suffice, when applied to the orbits of two electrons, to yield the spectrum of neutral helium. Yet the various pairs of orbits conforming to these quantum conditions, which have already been traced, have been shown (with vast expenditure of intellectual labor by some of the ablest mathematical physicists of our time) to entail energy-values for the Stationary States which are hopelessly incorrect.

For instance, one might assume that when the helium atom is in its Normal State, the two electrons are revolving in a common circular orbit about the nucleus, being at each instant located at opposite ends of a diameter; and that this permitted orbit is determined by the condition that the angular momentum of each electron, or perhaps that of both together, is $h/2\pi$. This seems an obvious generalization of the Quantum Conditions for hydrogen; but it yields a false energy-value for the Normal State; and there is nothing more to be said. Kemble and van Vleck demonstrated that no arrangement in which the two electrons are symmetrically placed relatively to a line through the nucleus entails the proper energy-value for the normal state. This still leaves open the possibility that the two electrons are unsymmetrically placed—a possibility which to some people seems repellent enough to be excluded. Born and Heisenberg calculated the energy-values corresponding to pairs of orbits, one of which lies far beyond the other at all points, and both of which are concordant with the Quantum Conditions. These ought to have agreed with the energy-values of the Stationary States which are remote from the normal state and near the state of the ionized

atom; but they did not. This result is commonly regarded as the strongest evidence for the belief that the Quantum Conditions valid for an atom with one electron are not valid for an atom with two.

The atom-model favored by Kramers, and hence presumptively by Bohr, to represent a helium atom in its normal state, involves two electrons moving in orbits which are not coplanar nor even plane. Planes tangent to the two surfaces upon which the orbits are traced intersect each other at 120° along a line passing through the nucleus, and the electrons pass simultaneously across this line at opposite crossing-points. These orbits conform to the Quantum Conditions; and the resultant of the angular momenta of the two electrons, which is the angular momentum of the entire atom, is equal to $h/2\pi$. This atom-model likewise fails to have the right energy-value for the normal state.

P. Interpretation of the Optical Spectra of Alkali-Metal Atoms

The alkali metals (lithium, sodium, potassium, rubidium and caesium) are elements of which the atoms are easily deprived of a single electron apiece; one electron of each atom is, as the phrase goes, exceptionally "loosely bound." Many facts combine to indicate this; for instance, each of these elements enters with violence into chemical combinations, and the compounds which each forms are such as to suggest that its atom yields up one electron to the atom or atoms which join with it. Again, when a salt of one of these metals is dissolved, the molecules split up and the atoms of the metal are left wandering around in the solvent minus one electron apiece, while the atoms of the other element each hold on to one captured electron. More definite yet is the direct evidence that the ionizing potentials of the alkali metals are lower than those of any other elements in the same rows of the Periodic Table, those of rubidium and caesium being altogether the smallest known. These alkali metals follow, in the Periodic Table, immediately after the five noble gases helium, neon, argon, xenon and krypton respectively. These gases are chemically all but absolutely inert, almost never entering into combinations. Their ionizing-potentials are higher than those of any other elements in their respective rows of the periodic table, and those of helium and neon are the greatest known. The atoms of each of the alkali metals are much larger than those of the preceding inert gas.

From all these facts the inference is drawn, that the atom of each inert gas consists of a nucleus and electrons, at least the outermost

ones of which are arranged in a peculiarly stable and symmetrical fashion (as for instance, in a group of eight at the corners of a cube, though this is by no means sure); while the atom of the next following alkali metal consists of just this sort of arrangement or "inert-gas shell," now to be known as the "residue" or "kernel," and of one additional electron now to be known as the "valence-electron," usually much farther away from the nucleus.

If to such an atom-model we apply the doctrine of Stationary States, we may infer that for each and every arrangement of the electrons in the residue, or (to use more general terms) for each condition of the residue, there is a whole system of Stationary States differing from one another only in that the valence-electron travels in different ones among a system of quantized orbits. These orbits we may suppose to conform to the quantum-conditions (2) and (3), at least until convincing evidence is brought to the contrary. Such in fact is the interpretation of the system of Stationary States, transitions between pairs of which are responsible for the "optical spectrum" of each alkali metal.

An electron at a very great distance from the kernel of such an atom will experience an attraction towards it, practically indistinguishable from the attraction which would be exerted by a single (hydrogen) nucleus of charge $+e$. One might say that the $(Z-1)$ electrons surrounding the nucleus of charge $+Ze$ effectively cancel a portion $+(Z-1)e$ of the nuclear charge; or to use a more common word, that they "screen" it. As the imagined distant electron moves inward towards the kernel, the screening will cease to be perfect. An effect should occur analogous to the "stray field" which penetrates the meshes of a grid; since the electrons of the kernel do not form a continuous shell of electricity enclosing the nucleus, the latter should make itself felt through the interstices, although this effect may be diminished by the swift motion of the electrons. All this is speculation of the wildest kind. The only deduction reasonably safe is this, that very far outside the kernel the field will be very nearly the inverse-square field due to a hydrogen nucleus of charge $+e$; very near to the kernel the field will be quite incalculable [1]; while in between the very-far-out and the very-near-in region, there will be an intermediate region, in which there may be some chance of finding an adequately approximate expression for the field. On the existence of such a region, in which such an approximation is good enough to be valuable,

[1] Unless it is violently simplified by some agency or restriction of which at present we know nothing.

rests all the present hope of achieving numerically valid theories in this division of atomic physics.

One agreement between this theory and certain data may be demonstrated without making any specific approximation. The farther away the valence-electron remains from the kernel, the more nearly identical with the field of a hydrogen nucleus is the field in which it revolves, the more nearly should it behave like the electron of a hydrogen atom. Consider for instance, in a hydrogen atom, the orbit which yields the Stationary State for which n the total quantum-number and k the azimuthal quantum-number are both equal to 5. This orbit is a circle of which the radius is 10^{-7} cm.; far larger than the radius of any inert-gas atom, presumably *a fortiori* far larger than the kernel of any alkali-metal atom. Were the valence-electron of such an atom to describe this circle, it would pass everywhere in a field very nearly like that of a hydrogen nucleus, and should very nearly conform to the quantum conditions for this field. It follows that an orbit drawn in the actual field, obeying the quantum-conditions $n=5$ and $k=5$, would be very nearly such a circle with very nearly the same energy-value. The inference is drawn that for high values of n and k, the Stationary States of an alkali-metal atom should be very nearly identical with those of hydrogen. These orbits which lie far out from the kernel of the alkali metal atom, or from the nucleus of the hydrogen atom, have small energy-values. It may therefore be said that if we tabulate the Stationary States of the two atoms in order of decreasing energy-value, then the farther along the two tabulations we go, the more nearly should the two systems of Stationary States coincide.

This is found to be true, under a limitation. The limitation is an important aid in interpreting the arrangement of the Stationary States. It will be recalled from the First Part of this article that the Stationary States of the sodium atom are arranged in several sequences (there illustrated as columns in Fig. 7) known as the *s*-sequence and the *p*-sequence and the *d*-sequence and the *f*-sequence and others; and to these sequences successive values 1,2,3,4 . . . of a symbol k were appended. One basis for this classification is that when it is made, the occurrence or non-occurrence of transitions between any pair of Stationary States, under normal conditions, can be determined by applying the "selection-rule" that only such transitions occur as involve a change of one unit in k. Now there are two reasons for supposing that the only transitions which can occur are those in which the Azimuthal Quantum-number of the valence-electron changes by one unit. Unfortunately it is not possible to introduce these two

reasons with all the necessary background without too long a stoppage of the main current of this argument.[2] I must therefore set it down as an assertion, that the selection-rule is deducible from the assumption that the value of k is the Azimuthal Quantum-number of the valence-electron; which thus is 1 for all the Stationary States of the *s*-sequence, 2 for each State belonging to the *p*-sequence, 3 for the *d*-sequence, and 4 for the *f*-sequence. The feature common to the various Stationary States of a sequence is, therefore, the Azimuthal Quantum-number of the valence-electron—if this atom-model is valid.

This being assured, the conclusion is drawn that, since k is higher for the *f*-terms than for the *d*-terms, higher for the *d*-terms than for the *p*-terms and higher for the *p*-terms than for the *s*-terms; since, therefore, the *f*-orbits are *ceteris paribus* more nearly circular than the *d*-orbits and less inclined to stretch down into the neighborhood of the kernel, the *d*-orbits more nearly circular than the *p*-orbits and the *p*-orbits more nearly circular than the *s*-orbits—therefore the approximation of the sodium terms to the hydrogen terms will be most nearly perfect for the *f* (and higher) sequences, less so for the *d*, less for the *p* and worst for the *s*-terms. This also is verified. It reinforces the opinion that the k-values assigned to the various sequences are actually their azimuthal quantum-numbers.

As the different States of a single sequence share a common Azimuthal Quantum-number, they must differ—supposing always that this atom-model is valid—in their Total Quantum-number. Consecutive States of a sequence presumably have consecutive values of the Total Quantum-number (although sometimes one meets with a break or a jolt in the continuity of a sequence, suggesting a departure from this rule). The meanings of the Total Quantum-number n and of the Azimuthal Quantum-number k for elliptical orbits are such, that n can never be less than k. Hence the value of n for the first Stationary State of the *s*-sequence may be unity, or greater; but the values of n for the first terms of the *p*-sequence, the *d*-sequence and the *f*-sequence may not be less than 2, 3, and 4, respectively.

Strange as it may seem, there is no perfectly satisfactory way of determining the value of n for all Stationary States. Generally it happens that the various States of an *f*-sequence, that of sodium for example, agree so closely with those States of hydrogen which form an n_4 sequence, that there is little hesitation in attaching to each of the *f*-States the same value of n as is borne by that State of the hydrogen atom which coincides with it so nearly. For instance,

[2] These being (*verbum sapienti*) the argument associated with the name of Rubinowicz, and the argument deduced from the Principle of Correspondence.

the first *f*-State of sodium has very nearly the same energy-value as the 4_4 State of hydrogen; the second *f*-State of sodium nearly coincides with the 5_4 State of hydrogen, and so forth along the sequence. Hence to the successive States of the *f*-sequence of the sodium atom one attaches with confidence the symbols 4_4, 5_4, 6_4, and so onward. In some cases this is practicable for the terms of the *d*-sequence also; but never for those of the *s*-sequence. The Stationary States of the *s*-sequence depart so far from those of hydrogen, that one cannot with any security conclude what values of the Total Quantum Number should be assigned to them. It used to be assumed that $n=1$ for the first term of the *s*-sequence and $n=2$ for the first term of the *p*-sequence, and the usual notation for the Stationary States reflects this supposition; which however is neither necessary nor probable.

All of the foregoing interpretations are based upon a theory of the alkali-metal atoms which may be summarized in this way: as the hydrogen atom is supposed to consist of a nucleus surrounded by an inverse-square field through which an electron travels always in one or another of certain orbits determined by quantum-conditions, so also the alkali-metal atom is supposed to consist of a kernel surrounded by a not-inverse-square field through which an electron travels always in one or another of certain orbits determined by identical quantum-conditions. As the Stationary States of the hydrogen atom correspond each to a certain orbit and are designated each by certain values of two quantities n and k, or for short by a symbol n_k indicating the features of that orbit, so also the Stationary States of the alkali-metal atom correspond each to a certain orbit and are designated each by a symbol n_k. For the hydrogen atom we recognize the proper n_k for each Stationary State because of the wonderful numerical agreement between Bohr's theory and the experimental values for the energy of each State. For the alkali-metal atom we can only guess the proper n_k for each Stationary States from indications of much lesser evidential value. We suppose, however, that $k=1,2,3,4$ for the various States of the *s*, *p*, *d* and *f* sequences, respectively; so that the *s*-sequence is like the n_1 sequence of hydrogen, the *p*-sequence like the n_2 sequence, and so on. Of the values of n we are moderately sure for the *f* and *d* sequences, quite uncertain for the terms of the *s* and *p* sequences.

One may now wonder whether it is possible to invent a central field, such that the orbits traced in it according to the quantum-conditions (2) and (3) would yield a series of energy-values agreeing with the observed energy-values of the Stationary States of (let me

say) the sodium atom. It takes a certain amount of faith to go about the business of designing such a central field; for the model imagined for the sodium atom involves ten electrons around the nucleus in addition to the one "valence" electron for the benefit of which the field is being devised; and one might expect these ten electrons to be rushing around the nucleus in uncoordinated and non-recurring paths, never at any two instants similarly placed and similarly moving, never at any two instants exerting the same influence upon the valence-electron. Yet the Stationary States of the sodium atom are as sharply defined as those of the hydrogen atom; and this may be thought to mean that the ten electrons of the kernel are constrained to a unity and a fixed relationship, like that of the members of a machine if not like that of the parts of a rigid body, which translates itself into an influence upon the valence-electron not unlike that of a central field.

At all events, several physicists working independently in various nations have taken the not inconsiderable trouble of devising central fields to fulfil the condition required; and they appear to have achieved a respectable success. It is not easy to decide what this success requires the rest of us to believe; perhaps it is formally possible to devise a central field to account for *any* set of Stationary States; I am not sure whether this question has been adequately examined. Some have felt confident enough to say that the results show which of the Stationary States correspond to orbits of the valence-electron which "penetrate into the kernel" and which to orbits that remain in all their circuit quite outside of the kernel. It is to be hoped that this problem will become clearer in the next few years. At this point I will add only, that the orbits traced for the valence-electron are rosette orbits in which the precession is very rapid, so that consecutive loops of a rosette are inclined at a considerable angle to one another. In the model for the hydrogen atom, the consecutive loops of a rosette orbit lie so close together as to be indistinguishable when drawn to scale on an ordinary sheet of paper (the separation between them was much exaggerated in Fig. 3 of the Second Part of this article). In these atom-models, the orbit looks rather as if it were drawn along the edges of the blades of an electric fan.

Q. Interpretation of the Optical Spectra of Other Elements

As soon as we step from the first column of the Periodic Table into the second, the obstacles to such a theory as we have hitherto tried to hold are gravely increased. There is evidence of several kinds

which seems to bear upon the arrangement of the electrons in the atoms; but some of it leads to conclusions opposite to those which the remainder suggests.

On the one hand, line-series are discernible in the spectra of elements in the second and the third columns of the Table, and even in those of some others; and from these line-series, systems of Stationary States are deduced which resemble those ascertained for the alkali-metal atoms; and it is natural to extend the same explanation from that case to these, supposing again that each atom consists of a nucleus and a certain number of electrons, all but one of which are tightly bound into a residue, around which the one remaining electron circulates in one or another of various quantized orbits.

On the other hand, the chemical behavior of these elements does not confirm this easy classification of the N electrons of an atom into $(N-1)$ very-tightly-bound electrons and one which is very loosely bound. Thus, the atoms of elements of the second and third columns of the Periodic Table—"alkaline-earth metals" and "earth metals," as they are called—when floating in water as the fragments of molecules of dissolved salts of these elements, are found to be deprived of two and of three electrons, respectively; and the composition of these salts is such as to suggest that the atoms of the other element or elements involved in them have annexed two or three electrons, respectively, from the alkaline-earth atom or from the earth-metal atom. These facts suggest rather that the N electrons of an alkaline-earth atom, or of an earth-metal atom, should be classified into $(N-2)$ or $(N-3)$ very-tightly-bound electrons and two or three which are loosely-bound, respectively. The very tightly bound electrons will be equal in number to, and presumably arranged like, the electrons of the atom of the next preceding inert gas. Henceforth I will reserve the word "kernel" for such a system, and the word "residue" for what is left behind when one electron is separated in fact or in imagination from the atom. Thus these two words will not mean the same thing except in special cases, such as those of the alkali-metal atoms.

Specifically, let us consider the four consecutive elements argon (inert gas, 18th element of the Periodic Table), potassium (alkali metal, 19th element), calcium (alkaline-earth metal, 20th element), and scandium (earth-metal, 21st element).

The evidence from chemistry and from electrolysis impels us to think that the argon atom consists of a nucleus surrounded by (eighteen) electrons tightly bound, in a stable and imperturbable arrangement; that the potassium atom consists of a kernel much like the argon atom, with one additional electron loosely bound and hence

generally far beyond; that the calcium atom consists of the same sort of kernel and two loosely-bound electrons, the scandium atom of the same sort of kernel and three outer electrons.

The Stationary States of the potassium atom have been interpreted as corresponding to various quantized orbits which a single outer electron describes around an unchanging residue; the lines of its spectrum have been attributed to leaps of this electron from one orbit to another, the residue remaining unaltered. There is nothing incompatible between this and the previous conception of the potassium atom.

The Stationary States of the calcium atom resemble, in their arrangement, those of the potassium atom sufficiently to make the same general sort of an explanation desirable,—to make it desirable to suppose that one electron is loosely-bound and remote from the nucleus, the other nineteen tightly-bound and near the nucleus; one loosely-held electron versus nineteen tightly-held ones. But the evidence from chemistry and electrolysis demands two loosely-held electrons versus eighteen tightly-held ones.

One might try to evade the dilemma by supposing that the calcium atom is a sort of three-stage construction, with eighteen electrons congregated in a kernel around the nucleus, a nineteenth far out by comparison with the nucleus, a twentieth far out by comparison with the nineteenth. For interpreting spectra, the residue of the atom would be the kernel or "inert-gas shell" and the nineteenth electron, the valence-electron would be the twentieth. For interpreting chemical data, the residue of the atom would be the inert-gas shell. This conception would rescue the interpretation of the calcium spectrum made after the fashion of the one just expounded for alkali-metal atoms. It would probably demand a larger atom, or a more shrunken kernel, than other data will allow.

Or one might suppose that the nineteenth and the twentieth electron are on the whole about equally remote from the nucleus, and yet it is possible for one of them to change over between any two of a vast system of quantized orbits without greatly affecting the other. There is certain evidence for this conception which I shall presently narrate.

Or one might suppose that the nineteenth and the twentieth electron are a system by themselves, and that each Stationary State corresponds to a particular configuration of this system, so that each line of the spectrum is attributed to a leap not of either electron separately but of both together. This idea seems to be gaining ground rapidly in dealing with atoms composed of a kernel and several outer electrons,

three or four or five or six or seven. The preceding notion might be brought under it as an especial case. If it is accepted the theory of atoms other than the alkali-metal atoms will inevitably be more complex than the theory mentioned for these in section P.

An interesting feature of some of these spectra discloses that the residue of the atom may exist in either of two distinct states. It will be recalled that the energy-values of the Stationary States have been measured from the state of the ionized atom, to which the energy-value zero is assigned. In this fundamental state, one electron and the residue of the atom are completely sundered; and the energy-value of any other Stationary State is the energy required to tear the electron completely out of the atom when the latter is initially in that Stationary State. This definition implies that the state attained when the electron is completely separated from the rest of the atom is determinate and unique. Such must be the case if the atom consists of an invariable nucleus and one electron, as in hydrogen; but if the atom contains several electrons, there is no *a priori* reason for excluding the possibility that there may be several "states of the ionized atom"; in each of these states one electron will be far away, but the residue will have as many different arrangements as there are different states. Extending this idea, one infers that there may be two or more distinct sets of Stationary States for certain elements, each set culminating in a different final configuration of the residue,—that is to say, of the ionized atom.

Several instances of atoms possessing two such distinct families of Stationary States are known; the most noted is probably that of neon, but I will describe the case of calcium, lately interpreted by Russell and Saunders and independently by Wentzel. Two families of terms "primed" and "unprimed," had been identified in the spectrum of this element, and important sequences of each could be followed sufficiently far to make the extrapolation to the limit not too daring. The limits were different, showing that the amount of energy required to separate an electron from an atom initially in its normal state had two values differing from one another by 1.72 equivalent volts. Consequently the residue may remain (it is not necessary to assume that it can long remain) in either of two States differing from one another (when the extra electron is far away) by this amount.

At this point a very significant numerical agreement enters upon the scene. The residue of the calcium atom, the *ionized-calcium* atom, has itself a spectrum which is known, and from which its system of Stationary States has been learned and mapped. Like the systems of Stationary States possessed by neutral atoms, this one includes

s, *p*, *d* and other sequences. The Normal State of the ionized-calcium atom belongs to the *s*-sequence; following the usual custom it may be called the (1, *s*) State. The State of next lowest energy-value, the "next-to-normal" State (so to speak) belongs to the *d*-sequence, and may be called the (3, *d*) State. The energy-difference between the (1, *s*) State and the (3, *d*) State is 1.69 volts. This agrees within the error of the experiments with that value 1.72 equivalent volts, which was found for the energy-difference between the two conditions, in either of which the residue of the calcium atom might be left after the twentieth electron is abstracted. This agreement shows that the extraction of the 20th electron from a calcium atom may leave the residue either in the (1, *s*) State or in the (3, *d*) State.

If now we remember that the ionized-calcium atom is comparable with the potassium atom (and with alkali-metal atoms generally) having as it does eighteen electrons very tightly bound as a kernel around the nucleus and one electron loosely held—then it is reasonable to use the same interpretation of its Stationary States as was expounded in Section P; and to suppose that when the ionized-calcium atom is in the (1, *s*) State that loosely-held electron is revolving in a certain n_1 orbit, and when the atom is in the (3, *d*) State the electron is revolving in a certain n_3 orbit. Thus the extraction of the 20th electron of the calcium atom may be supposed to leave the 19th electron sometimes in the one, sometimes in the other of these two orbits.

We may now inquire whether the 19th electron will always remain in its n_1 orbit, or in its n_3 orbit as the case may be, when the 20th electron reenters the atom, descending from one orbit to another. Here it is necessary to watch one's mental steps very closely; for one is liable to slip into the naive notion of a particular orbit, say for instance a 3_3 orbit, as a fixed and permanent railway-track around which the electron continually runs until something violent derails it. This could not be true unless (to take this special case) the 20th electron had no influence whatever upon the 19th. Were it so, every Stationary State of the one family would differ by the same amount, 1.69 equivalent volts, from the corresponding State of the other family. In fact, the energy-difference between corresponding States varies from one pair to another. This may well be simply because the approach of the 20th electron so alters the forces acting upon the 19th, that its orbit is changed both in geometry and in energy-value, while remaining still identified with the same values of its quantum-numbers. The experiments neither prove nor disprove this; it is commonly accepted as true.

It is a very important fact that the atom may pass from a State of one family to a State of the other,—in terms of the model, that the 19th electron passes from its n_3 orbit to its n_1 orbit, and simultaneously the 20th electron makes some transition or other of its own. The emitted radiation contains the energy resulting from both changes simultaneously, fused together without any discrimination.

R. Building-up of Atoms by "Binding" of Successive Electrons

I next point out that the processes whereby the lines of an optical spectrum are emitted may be regarded, if this theory of the atom is valid, as stages in the gradual formation of an atom. Consider the hydrogen spectrum to begin with; each line is emitted as the atom passes from one Stationary State to another of lower energy-value, the state of least energy being the Normal State of the perfected atom and the state of greatest energy being the condition in which the atom-residue and its electron are torn apart. The various lines of the spectrum correspond to various partial steps along the path from the latter of these states to the former, to various stages of the formation of a hydrogen atom from two separated parts. The specific conception of each Stationary State as a definite orbit of the electron about a nucleus merely reinforces this way of envisaging the process. In the spectra of ionized helium and of neutral helium, we read the testimony of the gradual formation of a helium atom out of a nucleus and two electrons initially quite dissevered. The various lines of the ionized-helium spectrum correspond to different stages in the advance of an electron from the state of freedom to the state of most stable association with a nucleus of charge $2e$, or in Bohr's language, to different stages in the "binding" of an electron by a nucleus of charge $2e$. The various lines of the neutral-helium spectrum correspond to stages in the "binding" of a second electron by a system composed of a nucleus of charge $+ 2e$ and an electron already bound to it. Thus the two spectra of helium testify to two consecutive processes in the upbuilding of a helium atom out of its constituent parts.

The process of building up an atom, by successive adhesions of electrons to an incomplete electron-system surrounding a nucleus—that is to say, the process of building a system of Z electrons around a nucleus bearing the charge Ze, out of a system of $(Z\text{-}b)$ electrons surrounding the nucleus, by consecutively adding b electrons one after the other—evidently occurs very profusely in intense high-current high-voltage discharges in vapours, such as the condensed

spark and above all Millikan's "Vacuum Spark." To take instances from the work of Millikan and Bowen, Paschen, and Fowler: in the spectra of such discharges lines have been identified which belong to atoms for which $Z=14$ and b has the several values 1,2,3,4 (four stages in the building of a silicon atom); and to atoms for which $Z=10+b$ and b has the several values 1,2,3,4,5,6. Many of these spectra of multiply-ionized atoms have not yet been analyzed, but the work is proceeding rapidly. There is reason to hope that within a few years we shall be in possession of interpreted spectra not only of many systems of Z electrons about a nucleus of charge Ze, but also of many systems of fewer than Z electrons about nuclei of charge $+Ze$. This may be highly important, as I will try to show by an illustration. We will consider two consecutive elements of the periodic table; sodium ($Z=11$) and magnesium ($Z=12$).

A Mg atom is imagined as 12 electrons around a nucleus of charge $+12\ e$. It is formed when one electron joins itself to a Mg+ion, which is composed of 11 electrons about a nucleus of charge $+12\ e$. For this process a spectrum is emitted, the so-called arc spectrum of Mg or "MgI" spectrum, which is known and analyzed. It shows that the normal state of the Mg atom is an s-state (probably of total quantum-number 3). It is likewise a singlet-and-triplet-spectrum. The first of these facts is taken to mean that the valence-electron, or *twelfth electron* (the reader will see the reason for this usage, the electron in question being the last annexed out of the twelve) of the Mg atom moves in a 3_1 orbit. The second is taken to mean something or other about the residue of the atom, as will be shown in section S.

This residue of the atom is itself formed when one electron joins itself to a Mg++ion, which is a group of 10 electrons about a nucleus of charge $+12\ e$. In this process the so-called spark-spectrum of Mg, or "MgII" spectrum, is emitted. It is known and analyzed. It shows that the normal state of the Mg+ion is an s-state (probably of total quantum-number 3). It is a doublet spectrum. The first of these facts is taken to mean that the valence-electron or *eleventh electron* of the Mg+ion, moves in a 3_1 orbit. The second is taken to mean something or other about the residue of the Mg++ion.

A very interesting question now arises: is the Mg+ion actually the same as the residue of the Mg atom? In other words: when a 12th electron is added to the group of 11 electrons about a nucleus of charge $+12\ e$, is the group of eleven left unchanged? If so, we have knowledge about this group from two sources. The character of the MgI spectrum (the fact that it is a singlet-and-triplet spectrum) teaches something about the group, though what it is is far from

clear. The character of the MgII spectrum teaches something about the group, viz., that its eleventh electron moves in a 3_1 orbit. If these two groups are just the same, then the two independently acquired facts about them may be united into a precious correlation. As a matter of fact it is generally assumed that they are nearly if not quite the same. A valuable piece of evidence bearing upon precisely this point, although relating to a different element, was described in the foregoing section.

This suggests that it would be a most desirable achievement to produce the spectra due to groups of $(Z-b)$ electrons congregated about a nucleus of charge $+Ze$, for some value of Z (the higher the better) and all values of b from 0 to $(Z-1)$. Were this done we could almost lay claim to having witnessed the creation of an atom from fundamental particles common to all matter. We could not quite make this claim, since the nucleus of charge $+Ze$ would still remain characteristic of that one kind of atom alone; but we should have made a substantial approach to it. However, there is no immediate prospect of achieving this except for the cases $Z=1$ and $Z=2$ which have already been considered. Our inability to produce the spectrum expected for $Li++$ (i.e. for $Z=3$ and $b=2$) acts as a barrier against utterly tearing down the electron-structures of higher atoms so that they can rebuild themselves before our eyes from the foundations.

The next important question may be introduced in this fashion. Suppose that nothing were known about the spectrum called MgII, therefore nothing about the process of adding an eleventh electron to a group of ten around a nucleus of charge $12\,e$. Knowledge would still be available about the process of adding an eleventh electron to a group of ten about a nucleus of charge $11\,e$; for this is precisely the process which creates the neutral sodium atom out of the Na+ion, and results in the emission of the NaI spectrum or arc spectrum of sodium. This spectrum is a doublet spectrum, and it shows that the normal state of the sodium atom is an s-state, probably of total quantum number 3. This last fact is taken to mean that the eleventh electron in a group of eleven electrons about a nucleus of charge $+11\,e$, is revolving in a 3_1 orbit. Could we have assumed that therefore the eleventh electron, in a group of eleven electrons about a nucleus of charge $+12\,e$, is revolving in a 3_1 orbit? There is no *a priori* certainty of this: but the observations on the MgII spectrum, as we have seen, confirm it (and also that the residue of the Mg+ion is like the residue of the Na atom, in causing the next added electron to produce a spectrum of the doublet type).

Were this generally true we could say that each atom in the periodic table is like the residue of the next atom following it; and that the mth electron in the nth atom is revolving in the same sort of orbit as the outermost electron of the mth atom, for every value of n and for every value of m less than that value of n.

However, it is not always true. To take another specific instance, consider the two elements potassium ($Z=19$) and calcium ($Z=20$). The spectrum KI, which is due to a nineteenth electron joining a group of 18 about a nucleus of charge $+19\,e$, and the spectrum CaII, which is due to a nineteenth electron joining a group of 18 about a nucleus of charge $+20\,e$, are dissimilar. The dissimilarity is not quite so great as to affect the normal states of the two systems, K and Ca+, composed of nuclei of charge $19\,e$ and $20\,e$ each surrounded by 19 electrons; both have as normal state an s-state, apparently of total quantum-number 4; it is inferred that in each, the 19th electron revolves in a n_1 orbit. If we consider, however, the first of the d-states (to which the total quantum-number 3 is commonly assigned), we see that in the KI spectrum it has a much larger energy-value than the Normal State, while in the CaII spectrum it has nearly the same energy-value. A short leap of the imagination leads to the conclusion that if we could examine the spectrum produced by a 19th electron joining a group of 18 about a nucleus of charge $+21\,e$, the d-state in question would have a smaller energy-value than any s-state. In this case it would be the Normal State itself,[3] and we should say that the 19th electron, in a group of 19 surrounding a nucleus of charge Ze, revolves in a n_1 orbit if $Z=19$ or 20, but in a n_3 orbit if $Z=21$.

This system of 19 electrons around a nucleus of charge $21\,e$ is a doubly-ionized scandium atom, Sc++. Its spectrum has not been produced, so that the foregoing sentences are still somewhat speculative. What gives them value is the inference that scandium marks a sort of a breach in the regularity of the Periodic System. For most of the elements in the Periodic System, it can be said that the atom consists of a residue which is like the atom of the preceding element, and an additional electron; and that in its turn this atom resembles the residue of the atom of the element next following. To this the regular periodicity of the properties of the elements is ascribed. But when we reach an element of which the atom has a residue distinctly different from the atom of the foregoing element, then the regular variation of the physical and chemical properties is interrupted. Scandium, as a matter of fact, is the first of a group of

[3] In the First Part of this article the impression may have been left that the Normal State of every atom is an s-state. This is not true; in some known cases the Normal State is a p-state, in others an f-state.

elements, the intrusion of which into the Periodic Table brings about a disruption of the simplicity of its first three rows. There are other such intrusive groups of elements, notably the celebrated groups of the rare earths. It is supposed that wherever such a group commences, there the residue begins to vary from one atom to the next. The spectroscopic evidence is lacking; it is awaited with extreme interest.

The reader will very probably have seen one or more tables of the distribution of electrons in atoms; tables in which it is stated, for instance, that the atom of sodium contains two electrons moving in 1_1 orbits, four in 2_1 orbits, four in 2_2 orbits, and one in a 3_1 orbit; or more succinctly that it contains "two 1_1, four 2_1, four 2_2 and one 3_1 electron." Such tables are built by piecing together bits of evidence, some of which are such as I have described in this section, while others are inferences from X-ray spectra, magnetic properties, or observations of still other kinds. That they are still highly speculative is confirmed by the fact that they are continually being remodeled. If we could produce the spectra corresponding to all the stages of formation of an atom, we should be able to set up a tabulation more reliable than any yet put together. Even then, however, we should be confronted with the question whether the addition of a new electron to a kernel fundamentally alters the distribution of those already there.

Having considered the facts at such length in this section, we are entitled to consider the theory. In the coupled cases of hydrogen and ionized helium it was shown by experiment, and rendered plausible by theory, that the Stationary States of the element with one electron and a double charge on its nucleus correspond exactly to those of the element with one electron and a single charge on its nucleus, and are endowed with fourfold the energy of these latter. This conclusion can be extended to cover the case of a valence-electron circulating in an orbit at a great distance from a kernel composed of $(Z-b)$ electrons and a nucleus bearing the charge $+Ze$. The field due to the kernel will at great distances approximate the field due to a solitary nucleus bearing the charge be. We have seen already that when $b=1$ (so that the total charge on the nucleus balances the total charge of the electrons, valence-electron included) the Stationary States corresponding to orbits for which n and k are large coincide with Stationary States of hydrogen. It follows equally that when $b=2$, the Stationary States for which n and k are large have approximately fourfold the energy of stationary states of hydrogen, and coincide approximately with Stationary States of ionized helium. This is verified by experiment, and so are the corresponding conclusions for the cases $b=3$ and $b=4$.

8. Interpretation of Multiplets

Heretofore in the Third Part of this article I have repeated the procedure adopted in the First Part, simplifying the actual facts by writing as though the Stationary States of each atom were arranged in sequences of individual terms, each sequence being distinguished by a particular value of the Azimuthal Quantum Number. Here as there, it finally becomes necessary to concede the complexity of the facts, and recognize that the sequences in question are sequences not of individual terms, but of groups of terms. Thus for instance the sodium atom possesses a *p*-sequence, not of individual terms but of pairs of terms—a pair $2p_1$ and $2p_2$, then a pair $3p_1$ and $3p_2$, then a multitude of other pairs. For another instance, the mercury atom exhibits a *p*-sequence not of individual terms but of triads of terms—a triad $2p_1$ and $2p_2$ and $2p_3$, then a triad $3p_1$ and $3p_2$ and $3p_3$, and then a procession of other triads. These sequences are collected into systems: an *s*-sequence and a *p*-sequence and a *d*-sequence and several more constitute a system. There are singlet systems and doublet systems and triplet systems and systems of still higher *multiplicity*; and each kind of system is distinguished by a certain manner of grouping of the terms which form its various sequences. Noteworthy and peculiar laws govern these groupings; in a doublet system, for instance, the *s*-sequence consists of individual terms, but all the others consist of pairs of terms; in a quartet system, the *s*-sequence is made up of single terms, the *p*-sequence of triads of terms, the remaining sequences of groups of four terms each. From the First Part of this article I reprint a Table showing how the terms are grouped in systems of all multiplicities from the singlet to the octet. The numbers opposite the name of each system and under the letters of the various sequences show how many terms belong to each group in the various sequences of that system.

TABLE I

Name of System	*s*	*p*	*d*	*f*	*f'*	*f''*
Singlet	1	1	1	1	1	1
Doublet	1	2	2	2	2	2
Triplet	1	3	3	3	3	3
Quartet	1	3	4	4	4	4
Quintet	1	3	5	5	5	5
Sextet	1	3	5	6	6	6
Septet	1	3	5	7	7	7
Octet	1	3	5	7	8	8

Each atom possesses one or more such systems of Stationary States; and the particular types which an element displays depend in a definite and fairly clear manner upon the position of the element in the Periodic Table, being in fact one of the most distinctive of the periodically-varying qualities. Each atom with an even number of electrons exhibits systems which are all of odd multiplicity, and each atom with an odd number of electrons exhibits systems which are all of even multiplicity; thus magnesium, with twelve electrons, has a singlet system and a triplet system, while sodium and once-ionized magnesium, each with eleven electrons, have each a doublet system, and neon with ten has a singlet, a quintet and two triplet systems.[4] Remembering what was said about the consecutive binding of electrons, it will be noticed that these facts show a regular difference between the binding of the Nth electron when N is odd and the binding of the Nth electron when N is even. Otherwise expressed, they show that a kernel of N electrons treats an oncoming member in one or another of two distinctive ways, according as N is even or odd. The influence of magnetic fields on spectra likewise shows that this complexity of the Stationary States is a quality not negligible, but primary.

The features of the atom-model hitherto described must be supplemented with some new one if it is to cope with such facts as these. We have represented (for example) the sodium atom in its $2p$ state by a "valence-electron" cruising with angular momentum $2(h/2\pi)$ in an orbit around a "kernel" composed of ten electrons and a nucleus. But there are two such states instead of one; if the angular momentum of the valence-electron is to be equal to $2(h/2\pi)$ for each of these, some other not yet mentioned feature of the atom must discriminate the two. One might, of course, again proceed as we did in discussing the "primed terms," by assuming that the kernel of the atom is in one condition when the atom is in the $2p_1$ state, and in another slightly different condition when the atom is in the $2p_2$ state. This would probably entail as many different conditions of the kernel as there are pairs of terms in the sodium spectrum—a great number, and yet small in comparison with the multitude which would be required by other atoms; yet such may be the eventual theory. However, it is possible to construct for these facts an atom-model out of two revolving parts, whereby different Stationary States of a group are represented not by varying the condition of either part separately,

[4] Hydrogen and ionized helium are not included under this rule. Helium shows a singlet and a doublet system together, a combination which violates the rule as stated, unless the doublet system is really a triplet system in which two states of each triad are too close together to be distinguished.

but by varying the relative orientation of the two. Although this theory has not been harmonized with those which I have hitherto recited, it is competent in its own field; and for that reason I present it.

We will imagine that the atom is represented by a combination of two flywheels, two whirling objects, endowed each with angular momentum. These angular momenta are vectors, pointing along the directions of the axes of rotation of the respective flywheels, and having certain magnitudes. I will designate them temporarily as P_V and P_R, each symbol standing for a vector generally and also (when in an equation) for its magnitude. The angular momentum of the entire atom, which is necessarily constant in magnitude and in direction so long as the atom is left to itself, is the resultant of P_V and P_R; a vector, pointing along the direction of the so-called "invariable axis" of the atom. I designate it by P_A. The following equation shows the relation between the magnitudes of these three angular momenta and the angle Θ between the two first-named, the angle which describes the relative orientation of the axes of rotation of the two flywheels:

$$P_A^2 = P_V^2 + P_R^2 + 2P_V P_R \cos \Theta \tag{6}$$

Remembering the successes which in dealing with the spectrum of hydrogen have resulted from assuming that the angular momentum of the entire atom is constrained to take only such values as are integer multiples $Jh/2\pi$ of the quantity $h/2\pi$, we make the same assumption here. We further make the same assumption for each of the flywheels separately; the magnitudes of the angular momenta P_V and P_R are supposed to take only such values $Vh/2\pi$ and $Rh/2\pi$ as are integer multiples of the same quantity $h/2\pi$.[5] These particular assumptions, frankly, are foredoomed to failure; but the failure will be instructive.

Making all these assumptions together, we see that in effect we have laid constraints upon the angle Θ which measures the relative orientation of the two flywheels. For if P_V is an integer multiple of $h/2\pi$, and P_R is an integer multiple of $h/2\pi$, then P_A which is fully determined by equation (6) cannot be an integer multiple of $h/2\pi$ unless Θ is very specially adjusted. To illustrate this by an instance (which will be clearer if the reader will work it out with arrows on a sheet of paper): if P_V and P_R are each equal to the fundamental quantity $h/2\pi$, and if P_A must itself be an integer multiple of $h/2\pi$: then cos Θ must take only the values, $+1$, $-\frac{1}{2}$, -1, which yield the

[5] All that is actually being assumed is, that P_V and P_R and P_A are all integer multiples of a common unit; nothing in this section will indicate either $h/2\pi$ or any other value as the precise amount of that common unit.

values $0, 2\pi/3, \pi$ for Θ, which yield the values $2h/2\pi$, $h/2\pi$, 0 for P_A. Any other integer values for $P_A/(h/2\pi)$ are unattainable by any value of Θ whatsoever; any value of Θ not among these three would yield a value for P_A not an integer multiple of $h/2\pi$, which is contrary to the assumptions. Thus, the assumptions that the atom is a conjunction of two whirling parts, and that the atom altogether and each of its two parts separately whirl with angular momenta which are constrained to be integer multiples of a common factor—these assumptions lead to the conclusion that the relative inclination of the two revolving parts is constrained to take one or another of a strictly limited set of values.

This essentially is the model devised by Landé to account for the complexity of the Stationary States. The several Stationary States which form a group belonging to a sequence—in other words, which share a common value of n and a common value of k, like the $2p_1$ and $2p_2$ states of sodium or the $3d_1, 3d_2, 3d_3$ states of mercury—are supposed to resemble one another in this, that each of the whirling parts separately has the same angular momentum in every case; and to differ from one another in this, that in the several cases the two whirling parts are differently inclined to one another, so that the angular momentum of the entire atom differs from one state to the next. The different Stationary States which share common values of n and k are supposed to correspond to different orientations of the two parts of the atom and to different values of its angular momentum.

I will now no longer disguise the fact that these whirling parts are, or at any rate have been, supposed to be precisely the valence-electron and the residue. To the former we should therefore assign these values for the angular momentum P_V: the value $h/2\pi$ for every state belonging to an s-sequence, the value $2h/2\pi$ for every p-state, $3h/2\pi$ for every d-state, and so on. Then to the angular momentum P_R of the residue we should assign a suitable constant value; a "suitable" value in this case being such a one, as would yield the proper grouping of terms in the various sequences of the system which the atom under consideration is known to have. Thus, for an atom-model to represent sodium with its doublet system we require a value for the angular momentum of the residue, such as will yield one permitted orientation when the atom is in an s-state ($P_V = h/2\pi$), and two when it is in any state for which $P_V = kh/2\pi$ and k is any integer greater than unity.

No such value can be found. The value $P_R = h/2\pi$ will not do; for, as was shown in the illustrative instance a couple of pages back, it yields three permitted orientations when $P_V = h/2\pi$, and (as can easily be shown) three for each and every other value of P_V which

is an integer multiple of $h/2\pi$. Thus it would form an adequate model for a system of Stationary States in which every group of terms in every sequence was a triad; but this is not a doublet system, nor even a triplet system, nor any other observed system whatever. To make this long story short; it is impossible to simulate any of the eight groupings of terms set forth in the eight lines of Table I by assuming that P_V, P_R and P_A are all integer multiples of $h/2\pi$ (or of any other common factor).

It is in fact necessary to put P_V equal, not to $h/2\pi$ and to $2h/2\pi$ and to $3h/2\pi$, but to $\frac{1}{2}(h/2\pi)$ and to $\frac{3}{2}(h/2\pi)$ and to $\frac{5}{2}(h/2\pi)$, for the *s* and *p* and *d* states, respectively. This use of "half quantum numbers" makes it possible to produce an adequate model for an atom possessed of a doublet system, by assuming that the angular momentum P_R of its residue is always $h/2\pi$, and that its two whirling parts must always be so inclined to one another that the angular momentum of the entire atom is an integer multiple of $h/2\pi$.

For (to work out one example, and one only) when we make $P_R = h/2\pi$ and $P_V = \frac{1}{2}(h/2\pi)$, then the greatest possible resultant that can be obtained by combining these vectorially is $\frac{3}{2}(h/2\pi)$ and the least possible one is $\frac{1}{2}(h/2\pi)$; these two extreme values being attained when the two component vectors are parallel and when they are anti-parallel,[6] respectively. If we permit for the resultant only such values as are integer multiples of $h/2\pi$, then there is only *one* that is permitted: the value $h/2\pi$—for this is the only such value lying within the possible range. Next, put $P_R = h/2\pi$ and $P_V = \frac{3}{2}(h/2\pi)$. All possible values of the resultant lie between $\frac{5}{2}(h/2\pi)$ and $\frac{1}{2}(h/2\pi)$; within this range there are *two* of the integer multiples of $h/2\pi$ which are the sole permitted ones. Next, put $P_R = h/2\pi$ and $P_V = \frac{5}{2}(h/2\pi)$. All possible values of the resultant lie between $\frac{7}{2}(h/2\pi)$ and $\frac{3}{2}(h/2\pi)$, and this range again includes *two* permitted values. Thus the model describes properly the grouping of the terms in a doublet system. I leave it to the reader to show that by putting $P_R = 2h/2\pi$, or $3h/2\pi$, or $4h/2\pi$, and treating P_V as in this foregoing case, he can reproduce the groupings of terms in quartet, or sextet, or octet systems, respectively, as Table I, describes them.[7]

A more drastic use of "half quantum numbers" is required to obtain an adequate model for atoms showing singlet and triplet and other

[6] This convenient word is used in German to describe vectors pointing in opposite senses along the same direction.

[7] The diagrams with arrows, offered by Sommerfeld in the fourth edition of his classic book, are very helpful in studying these models. Incidentally Sommerfeld's alternative way of arriving at the groupings of multiplet terms by compounding vectors is instructive.

systems of odd multiplicity. Thus to produce a singlet system it is necessary to put $P_R = \frac{1}{2}(h/2\pi)$ always; to set P_V equal to $\frac{1}{2}(h/2\pi)$ for all S-states, to $\frac{3}{2}(h/2\pi)$ for all P-states, and so forth; and to suppose that the two whirling parts of the atom are constrained to take only such relative orientations as yield values for P_A, the angular momentum of the entire atom, which are odd integer multiples of $\frac{1}{2}(h/2\pi)$. It is easy to see that there is but one such orientation for an s-state, one for a p-state, and one for any other kind of state. To produce a triplet, or a quintet, or a septet system, it is necessary to put $P_R = \frac{3}{2}(h/2\pi)$, or $\frac{5}{2}(h/2\pi)$, or $\frac{7}{2}(h/2\pi)$, respectively; and to retain the just-stated assumptions about P_V and P_A.

The question whether these models have any intrinsic truthfulness has now become acute. If there is any doctrine in contemporary atomic theory which appears to be multiply tested and approved, it is surely the doctrine that the angular momentum of the valence-electron is always an integer multiple of $h/2\pi$. Yet in this passage I have spoken as if this principle had been indifferently and casually discarded, and replaced by a new principle to the effect that the angular momentum of the valence-electron is always an odd-integer multiple of $\frac{1}{2}\ h/2\pi$. It is hard to evade or mitigate this arrant contradiction.

A way out may possibly be found by suggesting that the partition of an atom into "residue" and "valence-electron," while appropriate when calculating energy-values by the method mentioned in Section P, is not appropriate in this instance; that the two whirling parts of the atom are respectively a system composed of a part of the residue, and a system composed of the rest of the residue and the valence-electron. This seems most admissible for such an atom as magnesium, consisting as it is supposed of what I have called a "kernel," and two additional electrons outside. The two whirling parts may be the kernel rotating as a unit, and the pair of outer electrons also rotating as a unit. It may be profitable to push the analysis even further, and to consider the two outer electrons each as an entity possessed of angular momentum, their two angular momenta combining with one another in such a fashion as I have lately described for the two parts of the atom; this resultant angular momentum of the two may then figure as the P_V employed in constructing the atom-model. There are decided possibilities in this way of thinking; but it is doubtful whether the difficulty about half-quantum-numbers can ever quite be removed.[8]

[8] An unfortunate feature of Landé's model in its original form is that it requires us to believe that the residue of each atom is different from the completed preceding atom. For instance since Mg has a singlet and a triplet system, its residue must have sometimes $P_R = \frac{1}{2}(h/2\pi)$ and sometimes $P_R = \frac{3}{2}\ h/2\pi$; whereas for the Na atom in its normal state $P_A = h/2\pi$ by the theory.

It may be recalled from the First Part of this article that the different Stationary States of a group, sharing a common value of n and a common value of k, are distinguished from one another by having different values of a numeral which was designated by j and called the Inner Quantum-number. It was so chosen that the only transitions which occur are those in which j change by one unit, or not at all; while transitions between two states, in each of which $j=0$, are likewise missing. This numeral is correlated with the angular momentum P_A of the entire atom, in the theory here outlined. For systems of even multiplicity, P_A is equal to $jh/2\pi$; for systems of odd multiplicity, P_A is equal to $(j+\frac{1}{2})\,h/2\pi$.

The various Stationary States of a group differ slightly in energy—otherwise, of course, they would never have been discerned. The energy-value of an atom must be conceived therefore as depending not merely upon n and k, not merely on the rates at which the two whirling parts are separately spinning, but likewise upon their mutual orientation and hence upon j.[1] In this theory, the dependence of energy upon orientation must be postulated outright. We shall presently meet with a case in which the dependence of energy upon orientation can be foreseen, even in detail.

It appears from all these speculations, that a transition between two Stationary States is no longer to be conceived merely as a simple leap of an electron from one geometrically-definite orbit into another. A leap is indeed supposed to occur, but it is accompanied by a turning-inward or a turning-outward of the axes of rotation of the two spinning parts of the atom. The radiation which comes forth is a joint product of these two processes, in which however no features of either separately appear; only the net change in the energy of the atom, the algebraic sum of the energy-changes due to each process separately, is radiated as a single fused unit. Nature does not make the separation which our imaginations make.

T. Magnetic Properties of Atoms

Having used an orientation-theory to interpret the complexity of the Stationary States, we will now consider an orientation-theory developed to account for the effect of a magnetic field upon the Stationary States. There, it was supposed that the various States belonging to a single group are distinguished by various orientations of two spinning portions of an atom, relatively to one another. Here, it will be supposed that the various States which replace each individual State, when a magnetic field acts upon the atom, are distinguished

by various orientations of the spinning atom relatively to the field. It will presently be seen that the evidence for the orientation-theory is much more abundant and more nearly direct, in this case of magnetically-excited Stationary States, than in that former case of multiplets. This case in fact was the earliest to which an orientation-theory was applied; but for it, some quite different form of theory might have been developed for multiplets. Even here the data and the theory are not entirely concordant; but the concordance is so extensive, that the discord is sharply localized and identifiable.

From the Second Part of this article (Section L) I quote the principle that an electron (of mass μ) revolving in an orbit with angular momentum P is equivalent to a magnet of which the magnetic moment M is proportional to P, being

$$M = eP/2\mu c \tag{7}$$

Both P and M are vectors normal to the plane of the orbit and hence parallel to each other. If several electrons are revolving in divers plane orbits about the same nucleus, their separate angular momenta may be summed vectorially into a vector which is the angular momentum of the entire system, and their separate magnetic moments may likewise be summed vectorially into a vector which is the magnetic moment of the entire system; and these two summation-vectors will be parallel to one another, and related by the foregoing equation. Hence a rigidly-connected revolving framework of electrons—if such a thing there be—may be treated like a single electron, insofar as the ratio of magnetic moment to angular momentum is concerned. Wherever in the course of this article we have envisaged electrons, kernels, or atoms revolving with angular momenta prescribed as integer multiples of $h/2\pi$ or of $\frac{1}{2}\, h/2\pi$, we might have imagined these as magnets with magnetic moments prescribed as integer multiples of $eh/4\pi\mu c$ or of $\frac{1}{2}\, eh/4\pi\mu c$.[9] This is not necessary; though the relation between angular momentum and magnetic moment is derived directly from an equation valid for perceptible electric currents, it might not be true for individual electrons. Nevertheless we shall arrive at striking results, by supposing that it is.

When a magnetic field is applied to a multitude of radiating atoms, most of the lines of their spectrum are replaced by groups of several lines each, or "split up" into several components, as the phrase is. This signifies that each of the Stationary States of each atom is apparently replaced by several. One may infer that when an atom is

[9] The quantity $eh/4\pi\mu c$, the presumptive magnetic moment of an electron circulating in an orbit of angular momentum $h/2\pi$, is known as the *Bohr magneton*.

introduced into a magnetic field, each of its Stationary States is modified into one or another of several new States, differentiated from one another and from the original State to a small but appreciable extent. This might arise from some distortion or internal alteration of the atom by the field; and it will probably be necessary to adopt this view in some cases. But there is also a simpler effect which the magnetic field may have upon the apparent energy-values of the Stationary States, an effect not involving any deformation of the atom by the field—to wit, an orientation-effect similar to that which was assumed to account for multiplets. This we proceed to examine.

If an atom which is a magnet is floating in a magnetic field, it experiences a torque which tends to orient it parallel with the field. By saying that an atom is parallel or oblique to the field, I mean that the magnetic moment of the atom and therefore also its angular momentum, are directed parallel or obliquely to the field; and this usage will be maintained. Owing to this torque it is endowed with energy due to the field, in addition to its own intrinsic energy; this additional energy, which depends upon the inclination of the atom to the field, I shall call its *extra magnetic energy*. If the atom turns in the field, the amount of its extra magnetic energy changes; and if its magnetic moment suddenly changes, its extra magnetic energy also changes unless it simultaneously turns by just the right amount to compensate the change. If at the moment of passing over from one of its Stationary States to another, its inclination or its magnetic moment or both are changed; the amount of magnetic energy which it gains or loses will be added (or subtracted, as the case may be) to the amount of energy which it gains or loses because of the transition. The frequency of the radiation sent out or taken in by the atom will be equal to $1/h$ times the sum of two energy-changes of distinct kinds—not, as in the absence of magnetic field, to $1/h$ times the energy-difference between the two Stationary States alone. Thus the effect of a magnetic field upon spectrum lines might be ascribed, not to any deformation of the atom by the field, but to changes in the orientations or in the magnetic moments of the atoms occurring at the instants when they make their transitions. The question for us now is, whether the actual details of the observed effect can be interpreted in this manner.

Expressing the foregoing statements in formulae, in which M denotes the magnetic moment of an atom, H the magnetic field, and α the inclination of the atom to the field, we have for the torque which the field exerts upon the atom

$$T = MH \sin \alpha \qquad (8)$$

and for the "extra magnetic energy" of the atom due to the field

$$\Delta U = -MH \cos \alpha \tag{9}$$

In this last expression it is tacitly assumed that the extra magnetic energy is zero when the atom is oriented crosswise (at right angles) to the field. This is not an arbitrary, but a quite essential convention, justified from the atom-model.[10] Suppose now that the atom passes between two stationary states S' and S'', in which its internal energy, its magnetic moment and its inclination are denoted by U', M', α' and U'', M'' and α'', respectively. Were there no magnetic field, the frequency radiated would be

$$\nu_o = (U'' - U')/h \tag{10}$$

but owing to the field, the frequency radiated is

$$\nu_H = \nu_o + \Delta\nu = (U'' - U')/h + H\,(M' \cos \alpha' - M'' \cos \alpha'')/h \tag{11}$$

the term $\Delta\nu$ representing the displacement of the line by the field. The question is, whether this term can be equated to the observed displacements.

Consider the most tractable cases, those in which the so-called "normal Zeeman effect" is observed. In these cases a line of frequency ν_o is replaced by three, of which the frequencies are

$$\nu_o + \omega H,\ \nu_o,\ \nu_o - \omega H \tag{12}$$

corresponding to three values for the displacement $\Delta\nu$, which are expressed by

$$\Delta\nu = +\omega H,\ O,\ -\omega H \tag{13}$$

The quantity ω occurring in these expressions is a specific numerical constant. Comparing these with the expressions for $\Delta\nu$ in (11), we see that if our model is to be used to interpret the observations, then for the first of the three observed lines $M' \cos \alpha'$ must be greater than $M'' \cos \alpha''$ by the amount $h\,\omega$; for the second, $M' \cos \alpha'$ must be equal to $M'' \cos \alpha''$; for the third, $M' \cos \alpha'$ must be less than $M'' \cos \alpha''$ by the amount ωh.

Another way of putting these statements is, that in order to interpret the normal Zeeman effect in this manner it must be supposed that whenever a transition occurs, the projection of the magnetic moment

[10] The action of the magnetic field upon the revolving electron imparts to it an extra angular velocity about the direction of the field (the Larmor precession) and hence an extra kinetic energy which (to first order of approximation) is proportional to $-\cos \alpha$ and is zero when $\alpha = \pi/2$. This extra kinetic energy is the extra magnetic energy ΔU. It is profitable to derive the entire theory in this manner.

upon the direction of the field—for this is precisely what $M \cos \alpha$ is—either does not change at all or else changes by $\pm \omega h$. Sometimes it acts in the first of these ways, sometimes in the second, sometimes in the third; but never in any other.

This would result, if the behavior of the atom floating in the magnetic field were governed by two rules; *first*, that it may orient itself only in certain "permitted" directions such that $M \cos \alpha$, the projection of its magnetic moment upon the field-direction, assumes "permitted" values which are integer multiples of ωh; second, that whenever a transition occurs $M \cos \alpha$ either retains the value which it had initially, or else passes to one or the other of the two adjacent permitted values.

The first of these rules is stated more rigorously than is quite necessary; all that is required is to say that $M \cos \alpha$ is permitted to take only such values as belong to an equally-spaced series with intervals equal to ωh. The second rule is necessary.

The theory of the normal Zeeman effect is simply, that the atom does behave according to these rules. Radiation of the frequency ν_o occurs, either when the magnetic moment of the atom does not change and the atom does not turn, or when the magnetic moment changes and simultaneously the atom turns just so as to keep the projection of the magnetic moment on the field-direction constant. We shall later see that the latter of these two alternatives is the accepted one. It must be supposed that the atom, so to speak, capsizes when it emits the frequency ν_o while floating in a magnetic field; it flops over at the same moment as it passes from one stationary state to another. Radiation of the frequency $\nu_o + \Delta\nu$ or of the frequency $\nu_o - \Delta\nu$ occurs, as we shall see, when the magnetic moment of the atom changes; in some cases the atom capsizes during the process, in others it does not.

I now translate the foregoing rules from the language of magnetic moments to the language of angular momenta. The first rule is, that the atom may orient itself only in certain permitted directions such that $P \cos \alpha$, the projection of the angular momentum upon the direction of the magnetic field, assumes permitted values which are consecutively spaced at intervals of $(2\mu c/e)\omega h$.

Now it is a fact of experience, that in the cases of the normal Zeeman effect,

$$\omega = e/4\pi\mu c. \tag{14}$$

The rule therefore reads, that *the projection of the angular momentum of the atom upon the direction of the magnetic field is constrained to take certain permitted values, spaced at intervals of* $h/2\pi$.

We have supposed, in dealing with multiplets, sometimes that the angular momentum of the entire atom is constrained to take such values as are integer values of $h/2\pi$, and sometimes that it is constrained to take such values as are odd-integer multiples of $\frac{1}{2}$ $(h/2\pi)$.[11] In either case the permitted values of the angular momentum are spaced at equal intervals; and as the rule for the component of the angular momentum along the direction of the field bears the form which it does, we may well suppose that something in the order of nature constrains both the angular momentum and its projection to accept only values which form a sequence spaced always at that curious interval $h/2\pi$.

The total number of permitted orientations will obviously be limited by the actual magnitude P of the angular momentum. This being supposed always to be an integer multiple of $\frac{1}{2}$ $h/2\pi$, let it be written $P=2J(\frac{1}{2}\ h/2\pi)$. The permitted orientations are those which yield a series of values for the projection $P\cos\alpha$ spaced at intervals $h/2\pi$; let these be written

$$P\cos a = A_o,\ A_o - h/2\pi,\ A_o - 2h/2\pi,\ \ldots\ .\ A_o - mh/2\pi \qquad (15)$$

Nothing in the experiments thus far described gives the least notion of the value which should be assigned to A_o. All we know at present is that A_o cannot exceed P and that $(A_o - mh/2\pi)$ cannot be algebraically less than $-P$. Suppose in the first place that $A_o = P$; that is to say, that the atom may orient itself with its axis parallel to the magnetic field. Then the permitted orientations are those which yield this series of values of the projection:

$$\begin{aligned} P\cos\alpha = P,\ P - h/2\pi,\ P - 2h/2\pi,\ \ldots\ldots\ .\ P - mh/2\pi \qquad (16)\\ = 2J(\tfrac{1}{2}\ h/2\pi),\ (2J-1)(\tfrac{1}{2}\ h/2\pi),\ (2J-2)(\tfrac{1}{2}\ h/2\pi,\ \ldots\ ,\ 0 \end{aligned}$$

of which there are $(2J+1)$ in all. On the other hand, it may be that the atom is prevented from orienting itself parallel to the field; that the least permitted angle between the axis of the atom and the direction of the field is some angle yielding a projection A_o intermediate between P and $(P - h/2\pi)$. Then there are $2J$ permitted orientations altogether.

Summarizing the results of this last paragraph: if the angular momentum of the atom is an integer multiple $2J(\frac{1}{2}\ h/2\pi)$ of the fundamental unit $\frac{1}{2}$ $(h/2\pi)$, then according to the orientation theory

[11] It was remarked at the beginning of Section S that the evidence to be presented in that Section would support neither $h/2\pi$ nor any other particular numerical value for the fundamental unit of angular momentum; here, however, we have evidence for that value.

the atom is permitted to take either $(2J+1)$ or $2J$ distinct orientations in the field; the former number if it is, the latter if it is not permitted to set itself quite parallel to the field.

It will now be shown that these are by no means idle speculations; they bear directly upon certain facts accessible to observation. Before bringing up these facts it is necessary to abandon the policy of speaking exclusively about the "normal" Zeeman effect. This "normal" effect received its adjective because it agrees so excellently with the original theory devised years before quanta were dreamt of to explain the effect of magnetic field upon spectra. It is essentially because of this agreement that it is possible to develop the contemporary theory of the "normal" effect in a perfectly deductive fashion, using no new assumptions beyond those general ones of the orientation-theory. Most spectrum lines, however, are affected by a magnetic field in ways not compatible with the original theory; which is a consequence of the fact that the set of new Stationary States, whereby a magnetic field supplants each original Stationary State, in most cases does not conform to the laws previously set forth.

The laws to which it generally does conform were read from the spectra by Landé. The one feature in which the foregoing theory quite agrees with these laws is its prediction of the total number of Stationary States. A Stationary State for which the angular momentum of the atom is determined, by virtue of the theory of multiplets which filled the preceding section of this article, as being $2J\,(\frac{1}{2}h/2\pi)$, is actually found to be supplanted, when a magnetic field is impressed upon the atom, by $2J$ new Stationary States. This is in agreement with one of the two alternative predictions made a few paragraphs *supra*; to wit, with the prediction derived from the assumption that the atom cannot set itself quite parallel to the field. This agreement between the orientation-theory of multiplets and the orientation-theory of Zeeman effect considerably strengthens both.

The several Stationary States replacing a given original State are always equally spaced; but the spacing differs in amount from the value ωIIh or $eIIh/4\pi\mu c$ exhibited when the normal Zeeman effect occurs, and which we found it possible to deduce from the simple orientation-theory. The difference is this, that the actual spacing is a multiple of the value ωIIh by a factor g (generally lying between $\frac{1}{2}$ and 2) which depends upon the original State:

$$\Delta U = g\omega Hh \qquad (17)$$

The only ways hitherto used to accommodate the atom-model to this surprising and inconvenient factor g are tantamount to assuming that it enters into the relation between angular momentum M and magnetic moment P which was derived in Section L and written down here as equation (7); which relation is accordingly modified without discoverable reason into

$$M/P = g\, e/2\, \mu c \tag{18}$$

a very unsatisfying procedure. Lande found it possible to mitigate this process somewhat and at the same time produce a partial explanation of the formula quoted in the First Part of this article, whereby g is related to the factors K, R and J which, in the atom-model of the two whirling parts, measure the angular momenta of valence-electron and residue and entire atom respectively in terms of the common unit $h/2\pi$. This explanation involved the postulate that $g=1$ for the valence-electron and $g=2$ for the residue. It would therefore be necessary to justify, or to postulate without justification, not a multitude of such relations as (16) with a multitude of unforeseen values of g, but only a single such relation with a single unforeseen value of g. This is bad enough, but not so bad as if it were inevitable to assume that M/P may have a dozen different values in different cases.

It is now the occasion to recur to the extraordinary experiments of Gerlach which disclose the magnetic moments of individual atoms and verify the supposition that certain orientations are permitted and others inhibited. These experiments having already twice been mentioned in this series of articles, I shall spend no more space upon the method than is necessary to say that atoms in a narrow stream are sent flying across an intense magnetic field with a strong field gradient, by which they are drawn aside. Were the atoms tiny magnets oriented randomwise in all directions, the beam would be broadened into a fan; one edge of the fan would be the path of atoms oriented parallel to the field, the other edge would be the trajectory of atoms oriented anti-parallel to the field, while the space between the edges would be filled by the orbits of atoms pointed obliquely to the field. Actually Gerlach observed not the whole fan, but two or several separate diverging pencils of atoms, and between them vacant regions traversed by none. Certain orientations of atoms to field were unrepresented in the beam. Here for the first time there is direct evidence of discrete Stationary States, of quantum permissions and quantum inhibitions, not deduced from observations upon transi-

tions but drawn forthright from viewing atoms in equilibrium in their Normal States.

When from the diverging pencils one proceeds to determine the orientations of the atoms and their magnetic moments, one is confused by a possibility made clear in the foregoing pages, but unsuspected at the time when the first of these experiments were performed. I illustrate with the case of silver, the atoms of which flock into two diverging pencils with a quite vacant space between. At first it was naturally supposed that one pencil consists of atoms oriented parallel, the other of atoms oriented anti-parallel to the field. The deflections of the two pencils are such, that if this assumption is true then the numerical value of the magnetic moment of the silver atom agrees within the error of experiment with the value of $eh/4\pi\mu c$—agrees, therefore, with the notions that the angular momentum of the silver atom in its normal state is $h/2\pi$ and that the magnetic moment stands in the right and proper ratio $e/2\mu c$ to the angular momentum. The data were supposed to prove these notions. They also agree, however, with the suppositions that one pencil consists of atoms inclined at 60° to the field and the other of atoms inclined at 120°; in which case the magnetic moment of the silver atom would be $2eh/4\pi\mu c$, suggesting that the ratio of magnetic moment to angular momentum has twice the right and proper value. This inextricable tangling of the effect of orientation with the effect of magnetic moment makes it impracticable to deduce quite so much from the data as was at first thought possible; but plenty still remains. It is found that copper and gold behave like silver, as was to be expected from their positions in the Periodic Table. It is found that lead atoms, and (most surprising of all!) *iron* atoms are not deflected at all; so that either the magnetic moments of their parts balance one another completely, or else they all orient themselves crosswise to the field. Nickel, on the other hand, behaves as though its atoms had each a magnetic moment surpassing $2eh/4\pi\mu c$, while thallium responds as though that of its atoms were much less than $eh/4\pi\mu c$. Finally—lest the results seem too gratifying—it is found that bismuth atoms are deflected in a manner quite unforeseeable.

There is not time nor space to speak of the other method for determining the magnetic moments of atoms, by measuring the susceptibilities of great quantities of them in gases or solutions; but the measurements so made are also very helpful in determining the magnetic moments of various atoms and ions—various groupings, that is to say, of electrons around nuclei.[12] All such data are of im-

[12] For the status of such measurements in 1923, the first of this series of articles may be consulted. (This Journal, September, 1923.)

mense value, and no theory of the atom can be spared from the demand that it confront them and account for them.

U. Interpretation of X-Ray Spectra

By the term "X-ray" the reader may understand any radiation of which the frequency ν is so high, that the energy $h\nu$ of a single quantum is several times as great as the energy required to remove the most-easily-detached electron from an atom; greater, for instance, than 100 equivalent volts, so that the wavelength of the radiation is less than some 125 Angstrom units. The emission or the absorption of such radiation by an atom involves too great an energy-change to be attributed merely to a displacement of the valence-electron or even to combined displacements of the valence-electron and one or two others. This definition leaves a sort of "twilight zone" of radiations having frequencies somewhat but not much greater than $1/h$ times the ionizing-potential of an atom. Little is known about such radiations, and in this place they will not be considered.

The absorption of an X-ray quantum by an atom results in the extrusion of an electron from the atom. The emission of an X-ray quantum results from the passage of an electron within the residue of the atom from some original situation to the situation vacated by the extruded electron, or else into a situation vacated by an electron which itself has moved elsewhere within the atom. These statements contain the theory of the vast amount of data piled up by observations upon the emission and absorption of X-rays by matter.

To express the same statements rather differently: X-ray absorption-spectra and X-ray emission-spectra reveal, when analyzed for Stationary States in the manner used in analyzing optical spectra, that each atom with several or many electrons has a considerable number of Stationary States, distinguished from those we have heretofore discovered in that *each of them involves the absence of one electron from the atom*. Each of them is therefore strictly an "ionized-atom state," and yet there are several of them with extremely different energy-values. This signifies that the extraction of an electron from an atom rich in electrons may leave the residue in any one of several distinct conditions. These distinct conditions are the distinct Stationary States, transitions between which are responsible for X-ray spectra. Owing to this striking difference between the Stationary States hitherto described and these latter, I shall refer to these as the "X-ray Stationary States"—not that this name is a particularly good one.

Absorption of an X-ray quantum by an atom, then, results in a transition of the atom from its normal state to one of the "X-ray Stationary States." Emission of an X-ray quantum by an atom results from a transition of the atom from one into another of its "X-ray Stationary States"—a transition which begins in a condition in which the atom lacks one electron, and ends in another condition in which the atom lacks one electron. To take instances: radiation of an adequate frequency falling upon an atom in its normal state may put it over into an X-ray Stationary State known as the L_{II} state, an electron being extruded. Radiation of an adequate frequency (a higher frequency will be required) falling upon a similar atom in its normal state may put it over into another X-ray Stationary State of higher energy, known as the K state, an electron being extruded. The atom in the K state may then spontaneously pass over into the L_{II} state, emitting a radiation belonging to the X-ray emission-spectrum, its frequency being $1/h$ times the energy-difference between the K state and the L_{II} state. Later the atom may pass into still another state, such as the M_I state, by emitting radiation of some frequency ν'; the energy of the M_I state is therefore less than that of the L_{II} state by the amount $h\nu'$; calculating it thus, and then applying to normal atoms a stream of electrons or of quanta having energy just adequate to put them over into this M_I state, we find that this effect is duly produced.

Thus there is a thoroughgoing analogy between the genesis of optical spectra by transitions between the optical Stationary States, and the genesis of X-ray spectra by transitions between the X-ray Stationary States. The differences between the two kinds of spectra seem all to derive from the one fundamental difference between the two kinds of Stationary States; the former do not involve the absence of an electron from an atom, the latter do. In the optical region, for instance, we find that an atom in the normal state cannot be put into a particular excited state by any radiation except one of just the right frequency ν_o for which $h\nu_o$ is equal to the energy-difference between the normal state and the excited state in question. In the X-ray region, we find that an atom can be put into the K-state (for instance) by any radiation of frequency equal to or exceeding that critical frequency ν_o for which $h\nu_o$ is equal to the energy-difference between the normal state and the K state. This difference in behavior occurs because in the former case a quantum of radiation having frequency ν exceeding ν_o would have no place to put the left-over energy $h(\nu-\nu_o)$, whereas in the latter case this extra energy can be and is delivered over to the extracted electron as kinetic

energy, with which it flies away. This is known positively; for the extracted electrons can be observed, and their energy measured.

Spontaneous transitions from each of the X-ray Stationary States occur to some, but not to all, of the States of lesser energy. Some are evidently inhibited; and it is possible to lay down rules of selection, distinguishing those which are permitted. The complicated system of rules originally proposed has yielded place to a much simpler one, exactly similar to the one prevailing in the optical spectra. That is to say: it appears to be possible to assign to each of the X-ray Stationary States a certain value of a numeral k and a certain value of a numeral j, such that the only transitions which actually occur are those in which k changes by one unit and j either changes by one unit or does not change at all; while transitions between states in both of which $j=0$ are specially excluded. Furthermore, the various values of k and j thus assigned to the several X-ray Stationary States are identical with those assigned to the several States constituting a doublet system, such as we have met already in Section S, such as the sodium atom possesses; so that there is a complete correspondence between the system of X-ray Stationary States which every atom rich in electrons possesses, and the doublet system of optical Stationary States which only certain atoms possess. A part of this correspondence is expressed in the following Table:

TABLE II

Values of k:	1	1	2	2	1	2	2	3	3
Values of j:	1	1	1	2	1	1	2	2	3
Stationary States of									
Doublet system:	$1s$	$2s$	$2p_1$	$2p_2$	$3s$	$3p_1$	$3p_2$	$3d_1$	$3d_2$
X-ray system:	K	L_I	L_{II}	L_{III}	M_I	M_{II}	M_{III}	M_{IV}	M_V

No doubt the implications of this close correspondence are deep; but just what they are is not yet obvious.

The fact that the residue, left behind when an electron is extracted from an atom, may exist in any one of several distinct States, is quite naturally interpreted as meaning that the various electrons of the complete atom are variously situated, or revolving in various distinct orbits; so that the several X-ray Stationary States differ essentially in this, that differently-located electrons have been removed, leaving different places untenanted. This notion is easily combined with the idea that an atom is formed, or at all events behaves as though it had been formed, by successive self-annexations of electrons to a nucleus originally bare. Suppose that an atom is made by con-

secutive adhesions of electrons, each of which settles down into a peculiar orbit and remains there more or less unperturbed as the later comers immigrate one after the other into the system. May not then the process of X-ray absorption consist in a powerful intruding entity, electron or quantum, violently invading the interior regions of the atom and casting out one or another of the deeper-lying earlier-added electrons, while the later-added ones nearer or upon the frontier remain attached? May not X-ray emission consist in the passage of one of these latter electrons into the orbit formerly held by its predecessor, now unexpectedly reft away and its place left empty?

Although an affirmative answer to these questions involves a very literal and concrete conception of electron-orbits, most physicists make it, and would like to prove it. The chief difficulty lies in the fact that all information about X-ray Stationary States is primarily information, not about the prior condition of the electron which is gone, but about the final condition of the residue which is left behind.

The data show at all events that there are not nearly so many conditions of the residue, as there are electrons of the completed atom; from which it is fairly safe to conclude that the electrons of the atom are so arranged, that any one of several different electrons may be removed and the residue be left always in the same condition—therefore, that the electrons are arranged in groups, each electron being situated essentially like every other of its group. In discussing the formation of an atom by successive binding of electrons, it was remarked that several electrons may be bound in orbits each characterized by a common value of n and a common value of k. These two ideas may coincide; and great efforts are being made to bring them into entire coincidence. The evidence indicates, for example, that the first ten electrons bound to a nucleus are divided into four groups. Absence of an electron from one of these groups entails that the atom is in the K state; absence of an electron from the second, third, or fourth group brings it about that the atom is in the L_I, or L_{II}, or L_{III} state, respectively. So much the X-ray data do show rather definitely; although the actual number of electrons in each of the four groups cannot yet be deduced. If one could prove *a priori* that the first ten electrons annexed by a nucleus settle down into orbits of four distinct kinds, the achievement would be a great one. Intimations that something of this sort has been achieved are made every now and then; but it is difficult to tell whether the assertions which are made have been derived cogently from a principle or are inspired guesswork.

There is a remarkable numerical agreement in this field, the meaning

of which was until a couple of years ago regarded as perfectly distinct; but at this moment it is beclouded by one of the curious contradictions so abundant in the Theory of Atomic Structure. Briefly, the typical phenomenon is this: the differences between the energy-values of the K, L_{II} and L_{III} states agree notably well with what would be expected *if* the complete atom contains a few electrons moving in 1_1 orbits, a few in 2_1 orbits and a few in 2_2 orbits about the nucleus; and *if* the K state corresponds to absence of an electron of the first group, the L_{II} state to absence of one out of the second and the L_{III} state to absence of one out of the third. (The reason why calculations can be made for so indefinitely-phrased a model is this, that the field due to the highly charged nucleus of a massive atom should dominate over those of the individual electrons so that it does not make very much difference how many are supposed to be in each group.) The natural inference is, that the rest of the atom remains unchanged or little changed when any one of these electrons is extracted. In this case the Azimuthal Quantum-number of the residue should differ by one unit for the two states L_{II} and L_{III}. Consulting Table II, one finds that the quantity called k, which obeys the characteristic selection rule of the Azimuthal Quantum-number, is the same for L_{II} as for L_{III}. This is an illustration of the collisions between two sets of inferences which unsettle the supposedly firmest achievements of this theory.

Of the *theory of molecules*, a subject large enough for an article by itself, I can say here nothing more than that it attains some remarkable successes, achieved by and therefore fortifying some of the assumptions made in these pages; notably the assumption that Angular Momentum is a thing required in Nature to assume discrete values spaced at intervals of $h/2\pi$.

The final part of this long article has been very unlike the Second Part, in which an atom-model for the atoms of hydrogen and ionized-helium was constructed and endowed with certain fundamental qualities, so that it reproduced almost all of the relations of these atoms to radiation with a truly striking fidelity. This Third Part by contrast has been a thing of shreds and patches. Models for many atoms have been brought forth, but they have not been thoroughly adequate and they have not been concordant with one another. Some were designed with the same fundamental qualities as those given to the model for hydrogen; and scarcely more can be said for any of them, than that it does not positively clash with the properties of the element for which it is devised. Others were made competent to deal with a

certain limited set of facts (as the grouping of terms in multiplets) by giving them qualities gravely in discord with those attributed to the atom-model for hydrogen; and then they proved themselves surprisingly well able to account for isolated facts of quite a different sort (as the effect of magnetic fields upon atoms). The presentation in these pages is naturally very far from complete; had it been complete, it would have filled a book and not an article. But if it had been complete, the eventual impression would have been the same; an impression of confusion, yet of a confusion full of hope.

For the "Theory of Atomic Structure" is distinguished especially by this, that it is not one theory but a multitude of partial theories, each designed and competent to cover a limited family of the abounding data, each struggling to overlap and to absorb the others. It may be compared with a cross word-puzzle or a map-puzzle, in which the beginnings of a solution have been made in half-a-dozen corners and patches, while wide blank areas adjoin and separate them, and some of the partial solutions already entered upon the field may finally yield to others which can be unified into the perfected pattern. Or it may be compared with those maps of polar regions, in which here and there a properly-surveyed island or little strip of coastline emerges from the blankness of the unexplored realms, and some of them are certainly misplaced relatively to the others and will be shifted on the map when all the geography is at length made known. Or it may be compared with the state of a congealing metal, in which a multitude of little crystals have formed themselves about casual nuclei of crystallization; each is oriented in a different way, and when two of them grow into contact with one another they clash and cannot merge, they stand blocking and thwarting one another. It may be necessary to reliquefy them all and make a new attempt to change the formless mass into a single crystal.

Meanwhile the work is driven forward with the fervor of discovery and exploration, in this period which Russell finely called the "Heroic Age of Spectroscopy," and not of spectroscopy alone. Many, though not so many as are needed, are busy with determining the Stationary States by deciphering the rich and cryptic spectra of some among the numerous unstudied elements—enormously numerous, taking into account how many kinds of ionized atoms there are; and others with the assembling of new photographs of spectra made under the most varied sorts of excitation, with other aids to discriminating the lines; and others with the impressing of electric and magnetic fields upon radiating atoms; and others are engaged in measuring the intensities of lines. Yet other experimenters are determining the magnetic

moments of various atoms in all the possible ways. Some are seeking new phenomena which may result from Stationary States and from transitions, and occasionally they are rewarded with brilliant examples such as that vivid demonstration of the atom-magnets which Gerlach effected, or such as the passage of an atom from one State to another while it transfers the liberated energy to another particle directly and produces a chemical change. Others are finding the processes resulting from the Stationary States manifest on an unearthly scale within the stars.

The theorists likewise are at work with furious industry. Now and then a set of data hitherto rebellious is suddenly systematized, usually in a manner not quite concordant with the other theories holding other parts of the field. Attempts are made to unify one partial theory with another, usually unsuccessful. Sometimes an authoritative thinker, despondent over the continuing contradictions, tries to cut all the knots by declaring that one or another of the conflicting models is entirely fallacious, and that the numerical agreements on which it is founded are a delusion and a snare. Another is driven to concede that the conflicts are destined to endure forever, and accepts all of the partial theories as equally valid, or else paraphrases them in ingenious words which veil the contradictions, yet leaving these essentially unabated. Others, abandoning the general problem, have returned to the question of the hydrogen atom, and for this they are trying to rephrase or reshape the Quantum Conditions in a manner more satisfactory to themselves; sometimes with the aid of new and unfamiliar forms of mathematics, apparently expecting that when these become habitual to the human mind, the mystery of the Quantum Conditions will seem simple and clear. That, of course, always remains a possibility—that the human intellect will accustom itself so thoroughly to the new systems of ideas that they will cease to seem incoherent, as the human ear has so accustomed itself to the harmonic innovations of successive generations of musicians that the tones which seemed outrageous discords to the audiences of Beethoven now are to us monotonously sweet. To our minds the various divisions of the Atomic Theory are still discordant. It would not be fair to leave any other impression of this strange and fascinating theory; inchoate but full of promise, immature but gathering force, a fantastic assemblage of failures and successes; irreconcilable with all other theories, irreconcilable even with itself, and yet perhaps predestined to refashion all the science of physics in its own image.

Some Studies in Radio Broadcast Transmission[1]

By RALPH BOWN, DeLOSS K. MARTIN
and RALPH K. POTTER

SYNOPSIS: The paper is based on radio transmission tests from station 2XB in New York City to two outlying field stations. It is a detailed study of fading and distortion of radio signals under night time conditions in a particular region which may or may not be typical.

Night time fading tests using constant single frequencies and bands of frequencies in which the receiving observations were recorded by oscillograph show that the fading is selective. By selective fading it is meant that different frequencies do not fade together. From the regularity of the frequency relation between the frequencies which fade together it is concluded that the selective fading is caused by wave interference. The signals appear to reach the receiving point by at least two paths of different lengths. The paths change slowly with reference to each other so that at different times the component waves add or neutralize, going through these conditions progressively. The two major paths by which the interfering waves travel are calculated to have a difference in length of the order of 135 kilometers for the conditions of the tests. Since this difference is greater than the distance directly from transmitter to receiver it is assumed that one path at least must follow a circuitous route, probably reaching upward through higher atmospheric regions. Various theories to explain this are briefly reviewed.

The territory about one of the receiving test stations in Connecticut is found under day time conditions to be the seat of a gigantic fixed wave interference or diffraction pattern caused in part by the shadowing of a group of high buildings in New York City. The influence of this pattern on night time fading is discussed. It is considered a contributing but not the controlling effect.

Tests using transmission from an ordinary type of broadcasting transmitter show that such transmitters have a dynamic frequency instability or frequency modulation combined with the amplitude modulation. At night the wave interference effects which produce selective fading result in distortion of the signals when frequency modulation is present. It is shown that stabilizing the transmitter frequency eliminates this distortion. A theory explaining the action is given. The distortions predicted by the theory check with the actual distortions observed.

A discussion of ordinary modulated carrier transmission, carrier suppression, and single side band transmission is given in relation to selective fading. It is shown that the use of a carrier suppression system should reduce fading.

ONE of the factors which must be given increasing attention, if the technique of radio telephone broadcasting is to consolidate and continue its remarkable progress, is the mechanism of the transmission of radio signals through space. In many receiving situations the largest apparent defects present in the reproduced signal are those suffered not in the terminal apparatus but in transit through space, and in these cases better methods of utilizing the transmitting medium must precede any major betterment in overall results. In the present paper we are reporting some investigations in this field of radio transmission which have uncovered a number of interesting facts and have led to at least one conclusion which is of practical utility.

[1] Presented before the Institute of Radio Engineers, New York, Nov. 4, 1925

Night time transmission, which is the usual case in broadcasting, is in many places commonly marred by fading and sometimes by actual distortion of signals. Often these occur in certain areas not more distant from the transmitting station than other areas which enjoy freedom from such annoyance. Selecting a particular instance of these difficulties in an area near New York City which, in so far as can be judged at present, is probably a typical instance, we have subjected it to an intensive experimental study to determine what is the inherent nature of the troubles and if possible how they may be alleviated. In doing this it has been necessary to employ novel forms of tests especially fitted to bring out in a concrete way the phenomena being investigated.

To provide a suitable background for the subject we have started our discussion below with a brief recital of some of the things which a transmission medium is called upon to do. Following this we have described our tests, pointing out in what ways the existing media seem to fall short of doing these things and offering certain speculations as to the reasons for the shortcomings. In conclusion we have analyzed some practical problems in the light of this work.

Fundamental Considerations

As the radio art has progressed from spark telegraphy into continuous wave telegraphy and into high quality radio telephone broadcasting, increasing demands have been made on the transmission medium to deliver at the receiving point a true sample of what was put into it at the transmitting station. The requirements have grown in rigor because in telegraphy the end has been to develop increased reliability of communication at longer ranges and in telephony the medium is called upon to transmit a highly complex form of intelligence.

Of the requirements placed on the transmission medium by modern uses, those imposed by telephony are far more exacting than those for telegraphy. In telegraphy a single frequency, or at most a narrow band of frequencies sent out intermittently in accordance with a dot and dash code must reach the receiving station in such shape that it may be converted into audible sound for aural interpretation or into current pulses for the operation of relays or recording instruments. Leaving aside noise, the principal requirement is a sufficient freedom from fading so that signals can be interpreted or recorded without interruption. In radio telephony, as at present practiced in broadcasting, there is transmitted a modulated high-frequency wave comprising a relatively wide band of frequencies, usually at least 10 kilocycles.

Such a modulated high-frequency wave drawn out in the familiar graphical representation is a comparatively simple-looking thing, but analyzed into its elements and studied in detail it is revealed as being an intricate fabric of elemental waves so interwoven with each other that no one of them can be disturbed without changing in some degree the complexion of the whole. For perfect results the whole band must arrive at the receiver with an amplitude continuously proportional to that leaving the transmitter, or the inflections or expression of the speech or music will not be correctly reproduced. All the component frequencies within the band must be unchanged in their relative amplitudes lest the character of the sounds be altered. Even the relative phase relations of the various frequencies must be preserved or, as will be shown later, the interaction of the two side bands in the receiving detector will result in the partial loss of some of the frequency components.

It is not long since the time when radio was supposed to be the perfect medium for voice transmission it being presumed that since the ether of space (if there be such a thing) was substantially perfect in its electrical characteristics it must transmit frequency bands carrying telephone channels without distortion of any kind. This may be true theoretically of a pure ether but in fact, the ether used for radio communication is filled with a number of things ranging from gaseous ions down to the solid bed rock of the earth. It is rather to be expected that these will affect the progress of electromagnetic waves and we know from experience that they do. Diurnal variations of attenuation, fading, directional changes, dead spots and the like are already well known phenomena resulting from the complexity of our transmission media, although no entirely adequate explanations of their causes have been certainly established. One of the most recent manifestations of the effects of irregularities in transmission through space is in the distortion of the quality of telephone signals. This was perhaps first noticed in the use of short waves for broadcasting it being found that frequently the transmission was so distorted that after detection the signals such as speech and music were in severe cases almost unrecognizable.

Preliminary Investigations

For some time after quality distortion was recognized as a characteristic of existing short wave transmissions, it was thought that for the lower broadcasting frequencies at least, it was present only at night and at relatively very great distances from the transmitter. However,

careful observations demonstrated that there were points relatively near New York City where quality distortion from several broadcasting stations in the city was marked at night and in at least one case was detectable even in daytime. When station 2XB the Bell Telephone Laboratories' experimental station at 463 West Street, New York City, was used to transmit test signals, it was found that quality distortion could be observed in northern Westchester county and in southern Connecticut at distances of about 30 to 50 miles from the transmitter. Fading was also pronounced and it was noted as a significant fact that distortion was always accompanied by some fading although the reverse was not consistently true. In the course of these trials it was noticed that at a particular point near New Canaan, Connecticut, signals from 2XB were much weaker and more distorted than signals from 2XY, the experimental station of the American Telephone and Telegraph Company at 24 Walker Street, New York, even though the transmitter at 2XB was about ten times more powerful. Daylight field strength measurements at this point showed that the field strength of 2XB was only one-third that of 2XY. This led to the rather startling conclusion that there is a ratio of 100 to 1 in the power efficiency of transmission to that particular receiving point from these two transmitting stations in New York which are only about one mile apart.

In order to throw some light on this state of affairs a field strength survey was made by G. D. Gillett which resulted in the field strength contour map [1] here reproduced in Fig. 1. The contours on this map show that there is a series of long nearly parallel hills and valleys of field strength which, extrapolated, would converge in lower Manhattan and which extend out to the northeast as far as it was thought worth while to follow them. There has occurred to us no better explanation of this hitherto uncharted form of field strength distribution than that it is a gigantic wave interference pattern. A detailed discussion of this theory is given in another section of this paper.

The fixed pattern shown by Fig. 1 is definitely present only in the daytime but that it is fixed is attested by the fact that a second survey made about a year later checks with the original one quite closely. At night fading is pronounced in the area covered by the pattern and it is apparent that some other factors must enter. As a result of an endeavor to check up the pattern at night it was discovered that

[1] This map was prepared by Mr. Gillett using the methods discussed in a paper "Distribution of Radio Waves from Broadcasting Stations Over City Districts," by Ralph Bown and G. D. Gillett, I. R. E. Proc., Vol. 12, No. 4, p. 395—August, 1924.

quality distortion was, in general, most evident at places which were, by day, in the valleys of the field strength diagram and a point in one of these valleys near Stamford, Connecticut, was selected for the establishment of a temporary field test station. The interior of this station, which was in the empty hay-mow of a barn, is illustrated by the

Fig. 1—Radio contour map showing wave interference pattern

photograph, Fig. 2. At this place apparatus was set up to enable a study of the nature of the distortion in signals from 2XB. Many of the records discussed in succeeding paragraphs were taken at this Stamford field station. Others were taken near Riverhead, Long Island, which was also found to be well located for such work. Fig. 3

is an outline map showing the relative positions of these field receiving stations and the transmitting station.

The reason for settling down at a fixed point in this way was to attack the problem from a new angle. The field strength survey and aural observations had yielded much interesting information but did not appear at that time to shed a great deal of light on the quality distortion so it was decided to attempt, by an oscillographic

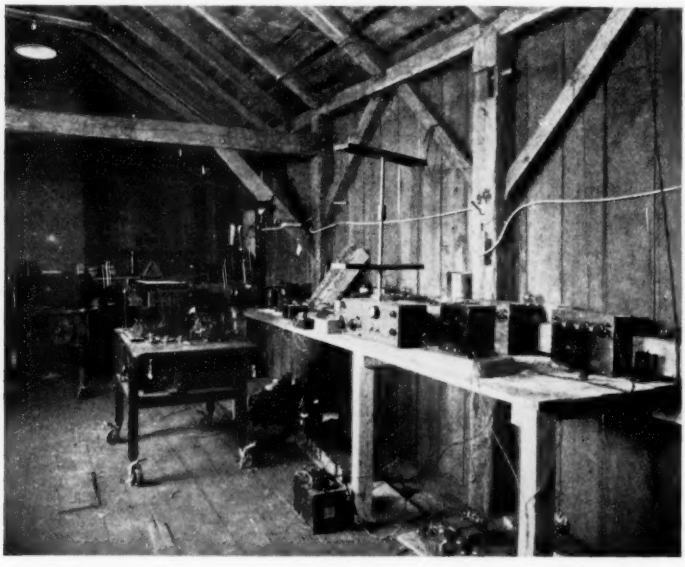


Fig. 2—Interior view of test station near Stamford, Conn.

study of received signals sent out under rigorously controlled conditions, to determine just what alterations these signals suffered in their journey through space.

In finding such distortions the ear is, of course, the primary testing instrument or indicator of trouble, for, if the trained ear is unable to detect anything wrong with a received signal in comparison with its original counterpart it is safe to say that nothing detrimental of importance has happened to it. But the ear is a poor quantitative indicator and furnishes no permanent or easily analyzed record of its observations. It is evident that if we are to study quantitatively the characteristics of radio transmission which give rise to quality distortion,

we must devise tests which will disclose changes, of whatever kind, in the relations between the various component frequencies of the transmitted band and furnish interpretable permanent records. In

Fig. 3—Outline map showing locations of transmitting station and receiving test stations

fact in the studies described herein a considerable portion of the job was to devise or perfect suitable methods of attack.

Single, Double, and Triple Frequency Tests

The variable factors in radio transmission which may be directly controlled are located at the transmitter and receiver. We have as yet no tangible means of controlling the transmitting medium, but it can be studied indirectly through the characteristics of the received signals. Obviously, it is desirable in the interest of simplicity to stabilize the apparatus variables to the extent that they may be idealized in considering observed results. Furthermore, at both the transmitter and receiver, it is desirable to make the antenna arrangements of the simplest form. For our work the normal antenna arrangement at station 2XB was used perforce since any important changes would have constituted a major operation. It is far from a simple arrangement, as shown in Fig. 4 which is an outline elevation and plan of the antenna and building at 463 West Street, New York City. Fortunately there are no buildings considerably higher than the antenna within a distance of several wave lengths.

At the receiving test stations both loop and vertical antenna were used; but in most of the experiments a simple vertical antenna was employed. It was constructed of brass tubing, 30 feet long, and guyed in a vertical position. A galvanized iron pipe 12 feet long was driven in the earth for a ground connection. The vertical receiving antenna projected through the roof of the test station building

at Riverhead, L. I., as shown in Fig. 5. The receiving antenna was not tuned but was connected to the radio receiver through fixed inductive coupling.

The carrier power in the transmitting antenna normally remains fairly constant, except for minor variations in voltage of the supply mains, and with a little care on the part of operating personnel, the

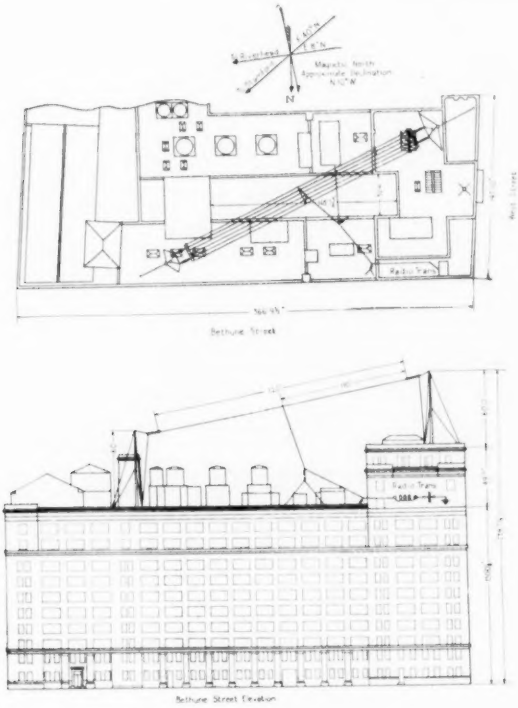


Fig. 4—Plan—elevation of the transmitting antenna

antenna current can be kept within the limits of a 1 per cent. variation, which is small compared with the signal fading usually experienced.

The stabilization of the frequency was of the greatest importance since in some of the tests it was desired to beat or heterodyne the signals down to audio frequencies and pass them through narrow band filters. To provide this stability engineers of the Bell Telephone Laboratories arranged the 5-kw. transmitter at station 2XB to obtain

its carrier frequency by amplification of the output of a 610-kc. piezo-electric crystal oscillator.

When desired some of the antenna current from the output of the transmitter was rectified and the resulting current was sent over a telephone line to the receiving station so that the frequency and wave

Fig. 5—Receiving test station near Riverhead, L. I. showing vertical antenna projecting through roof of building

form of the modulating signal could be seen and photographed at that point, thus guarding against any possible distortion in the transmitter and enabling a direct "before and after" comparison to be made. The telephone circuit was also used for communication between engineers at the two terminal stations.

At the receiving station double detection receivers and audio frequency amplifiers were employed. These did not have entirely "flat" transmission characteristics over the audio frequency band, but in most of the tests this was of no importance. In cases where it affected the results the making of necessary corrections was a simple matter. In tests involving beating the received signals down to audio frequencies through the agency of a local heterodyning frequency,

this was supplied from a shielded vacuum tube oscillator which on comparison with a standardized piezo-electric oscillator was found to possess the required stability. The double detection type receivers were used for no other reasons than their availability and their convenience for quantitative work. The beating down oscillator within the sets and the intermediate frequency step passed through in the sets by received signals do not figure in the following discussion of test methods but, of course, in each case the necessary set tuning

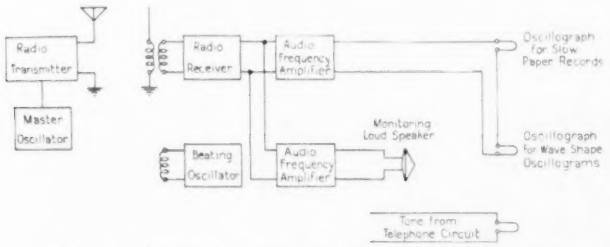


Fig. 6—Diagram of system used for single frequency tests

adjustments were made. To avoid confusion it is well to think of these receivers as being replaced by high frequency amplifiers and simple detectors since the local beating oscillator referred to in later pages is the separate shielded oscillator described above which is used to beat the signals down to audio frequencies.

In this work the moving coil type oscillograph was used throughout for the purpose of making photographic signal records. As indicated in Fig. 6 two oscillographs with elements connected in series were employed; one for the purpose of making a continuous record of the variation in the amplitude of the signal using a slow moving photographic paper tape and the other to obtain the wave shape of the signal by means of the usual high speed photographic film drum. An element of one oscillograph was also used at times to record on the film drum the wave shape of signals rectified at the transmitting antenna and sent over the telephone lines.

Fig. 7 is the interior view of the test station at Riverhead showing the general arrangement of the oscillographs and accessory apparatus. This oscillograph equipment formed about the only fixed portion of the apparatus, other portions being changed from time to time for different tests. These arrangements will be described later in connection with the records which they were used to obtain.

In considering these various records perhaps we had best look first at the simpler ones and then proceed in a more or less orderly

fashion to the more involved ones. The simplest records are fading records of the unmodulated carrier frequency of 610 kc. At the receiver the carrier was heterodyned with a local oscillator to produce a beat tone of about 250 cycles which was fed through amplifiers to the oscillograph elements.

A representative sample of the form of signal records made in the manner described above which show the variation in the amplitude of

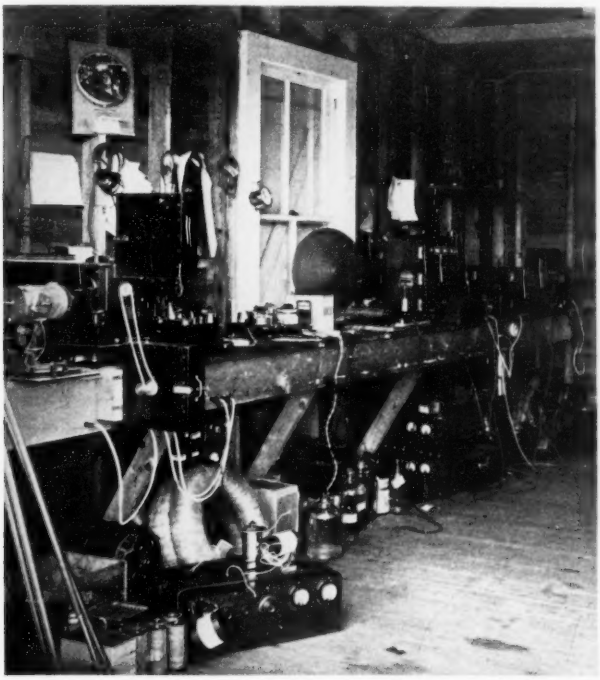


Fig. 7—Interior view of Riverhead testing station showing recording apparatus

the received carrier signal with time, is given in Fig. 8. It shows a typical fading record made at Stamford, Conn., May 16, 1925. The timing interval on strip 6 is 2.6 seconds.

The feed of the photographic paper tape through the oscillograph was varied somewhat during the course of the experiments but was generally in the range of 6 to 12 inches a minute. At this rate the record of an audible frequency signal is a shadow band of varying

width corresponding to twice the amplitude of the signal, as both the positive and negative half-cycles are recorded. It will be observed that the outer limits of the band corresponding to the peaks of the sine wave are darker than the center portion of the record. This is due to the fact that the rate of change of the movement of the light

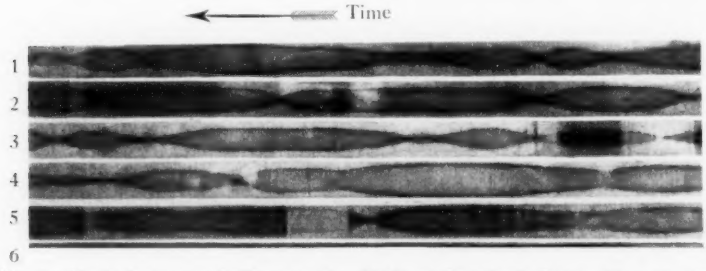


Fig. 8—Single-frequency fading record. Made at Stamford, Conn., May 16, 1924, 1:54 a.m. Timing marks, on strip 6, 2.5 seconds apart

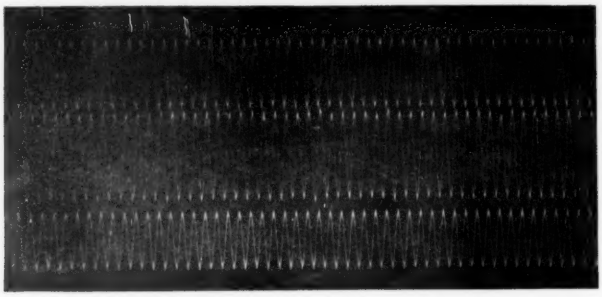


Fig. 9—Wave form of beat note signal for single-frequency test. Center trace signal from vertical antenna, upper and lower traces signals from loop antenna receivers

spot on the record is a minimum at the peak of the signal; hence, a greater quantity of light affects these portions of the record. This shading effect was very useful in the way it brought out changes in the distortion of the received signal. This is discussed fully in another section of the paper. The fuzzy irregular outline on portions of the records is caused by static and radio noise. The timing marks on the record allow a measurement of the time interval between points of minimum signal. Fig. 9 is a sample oscillogram of the wave shape of a beat note signal recorded by the method described above.

Marked changes in the fading cycle or time interval between points of minimum signal may occur within a period of a few minutes, and

from day to day there is often evidenced a modification of the general character and the recurrence of these changes. An example of this change in a short period of time is well illustrated by the oscillograms in Fig. 10. Strips 1, 2 and 3 form a continuous record starting at 1:52 a.m.; strips 4, 5 and 6 start at 2:16 a.m.; and strips 7, 8 and 9 start at 2:37 a.m. These are three sections of a continuous record

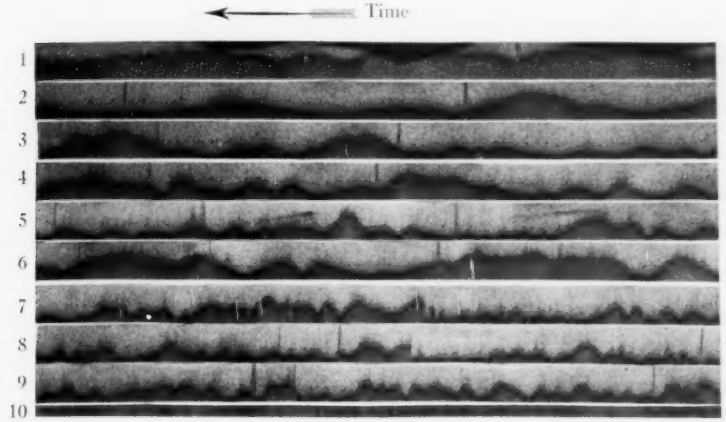


Fig. 10—Single-frequency fading record, showing variation in rapidity of fading, made at Riverhead, L. I., July 16, 1925, 1:52 a.m. Timing marks, on strip 10, 5 seconds apart

selected for the purpose of showing the decrease in the fading period, in a 45-minute interval. The timing interval on strip 10 which applies to these records is 5 seconds. In this particular record only half of the audio signal was recorded, the edge of the strip being the zero line.

These single frequency fading records do not offer very much to work on. There is, however, just enough suggestion of regularity about them to annoy one with the thought that perhaps they may follow some definite combination of periodicities and with this in mind we have taken sections of two different records and subjected them to a harmonic analysis.

So far we have been able to draw no more useful conclusions from such harmonic analyses than that the heterogeneous scattering of harmonic values is about what one would expect from the looks of the curves.

One significant thing about these oscillographic single frequency fading records is that they show no high speed fading of important

magnitudes. Occasionally one cycle of the beat tone will be somewhat upset by a sudden change in the amplitude but in general no changes which consistently distort the wave form were observed.

The slow fading may be considered as a modulation and on this basis the received signal is seen to be composed of the original constant carrier frequency accompanied by very narrow side bands occupying at best perhaps a fraction of a cycle.

The next progressive step in the radio transmission studies is naturally from a single frequency to two or more frequencies transmitted simultaneously. By the use of two crystal oscillators at the transmitter two separate and distinct radio frequency signals were transmitted simultaneously. These crystals were ground by the Bell Telephone Laboratories to oscillate at 610,000 cycles and 609,750 cycles. The amplitudes of these signals at the transmitter were controllable so that it was possible to make them equal, or one larger than the other, equivalent to the relative magnitudes usually found for the carrier and single side-band transmission case. Records were obtained of the variation of these radio signals, but none is reproduced here since the information shown by them can be just as easily obtained from the triple frequency records shown below.

Radio transmission on three frequencies is readily obtained by modulating the carrier with an audio frequency tone, and observing the three frequencies separately at the receiver.

If the modulating tone is

$$\sin (vt+\phi)$$

and the carrier signal

$$A \sin pt,$$

the transmitted signals are

$$+\frac{Aa}{2} \cos [(p+v)t+\phi] \qquad \text{(upper side band)}$$

$$+A \sin pt \qquad \text{(carrier)}$$

and

$$-\frac{Aa}{2} \cos [(p-v)t-\phi] \qquad \text{(lower side band).}$$

where a is a constant proportional to the percentage modulation.

These three frequencies are not merely a mathematical fiction but are physically existent as three separate waves bound together only at their point of origin.

To adequately record them separately by means of the oscillograph advantage was taken of the fact that a group of frequencies beaten

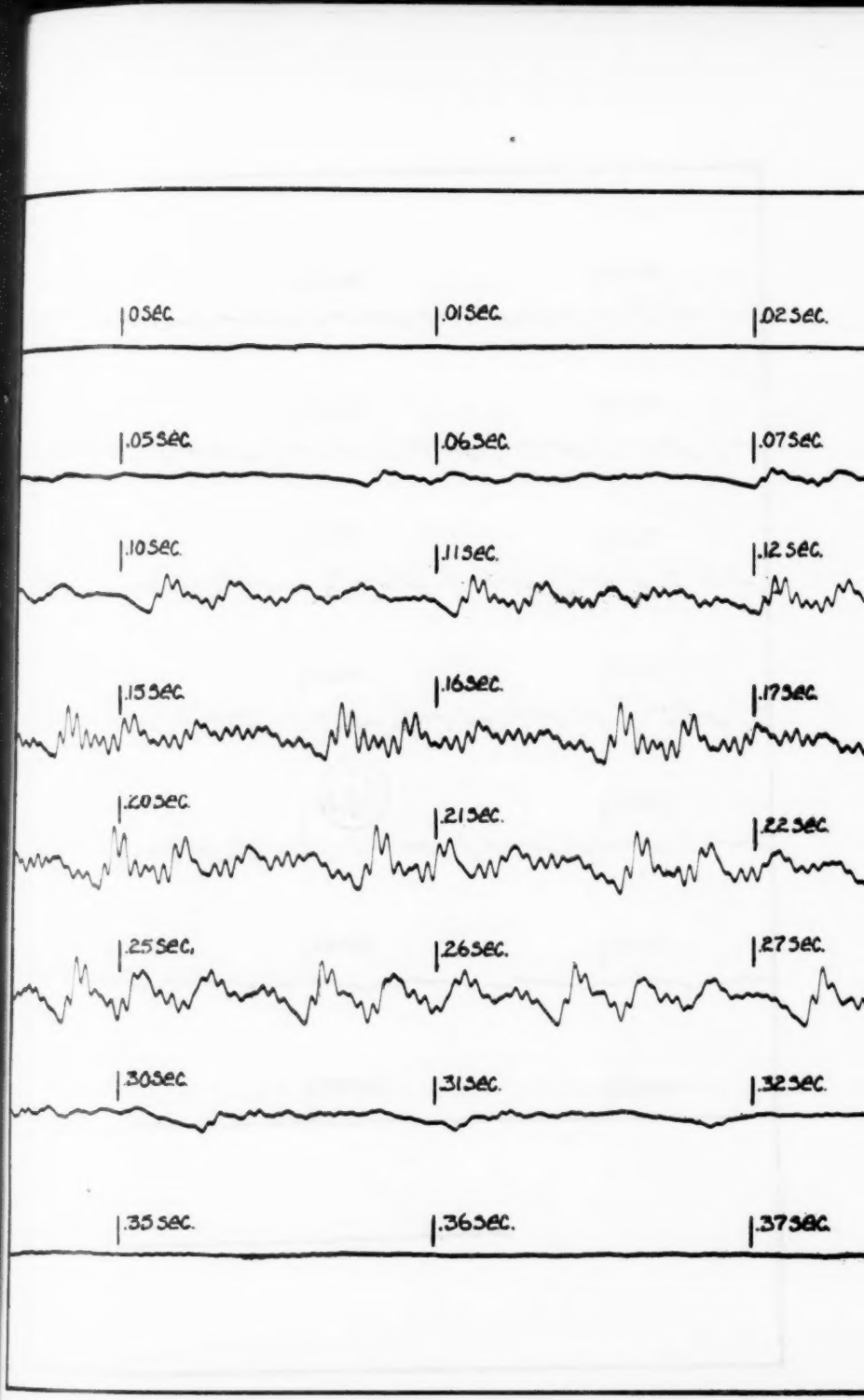
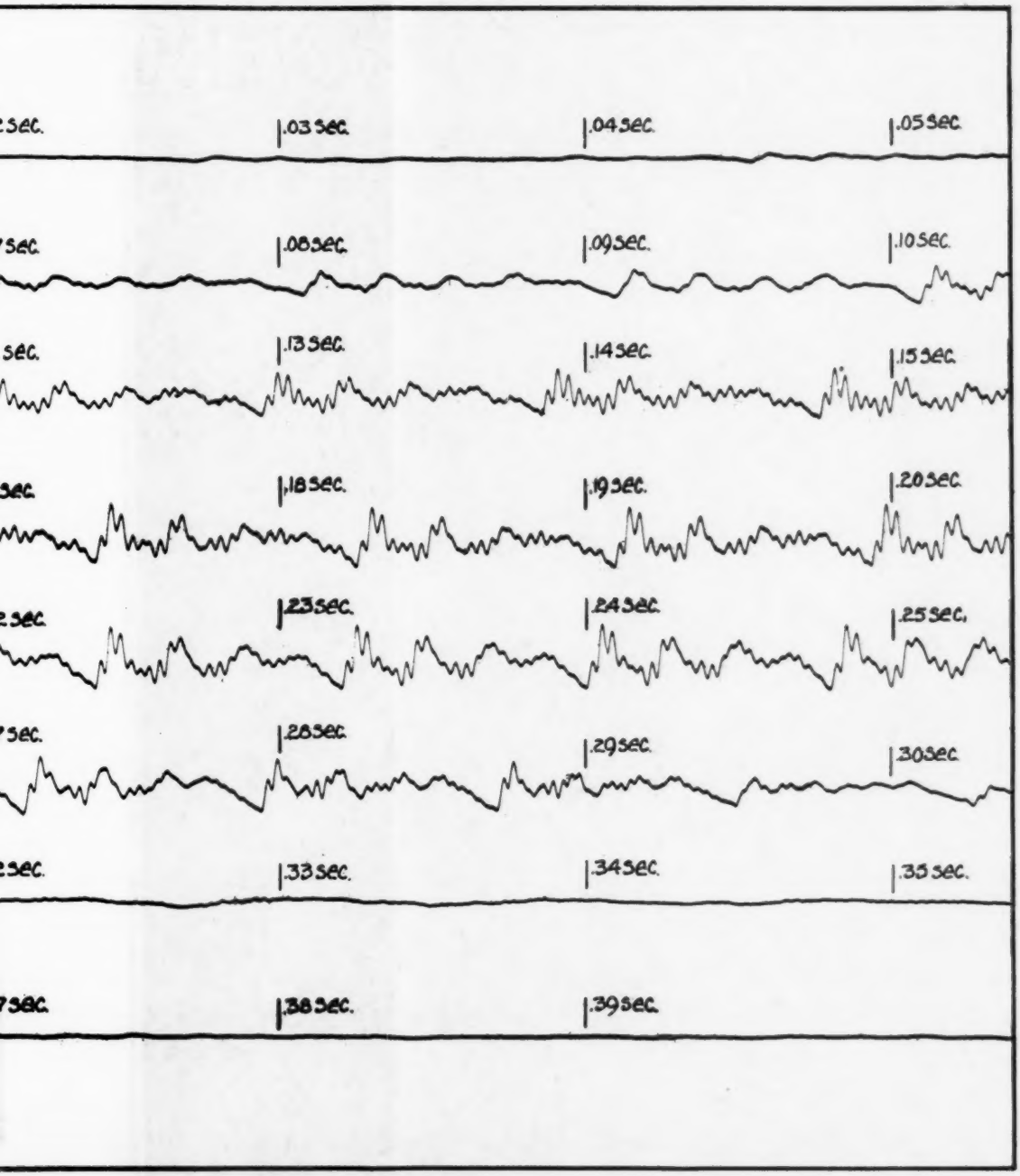


Plate No. 9—*u* as in *put*.



in *put*. Spoken by M.A.–Male, low-pitched (1)

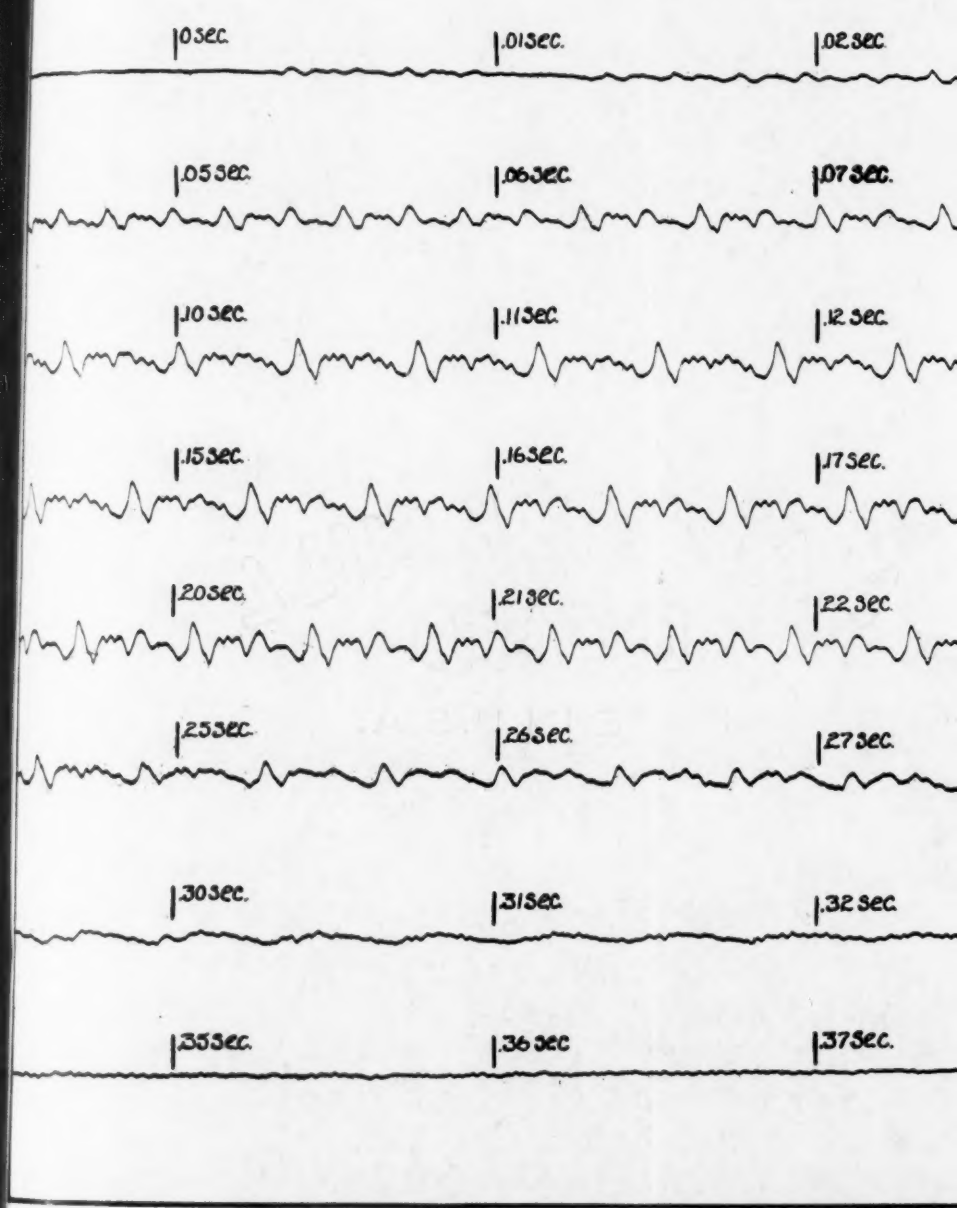
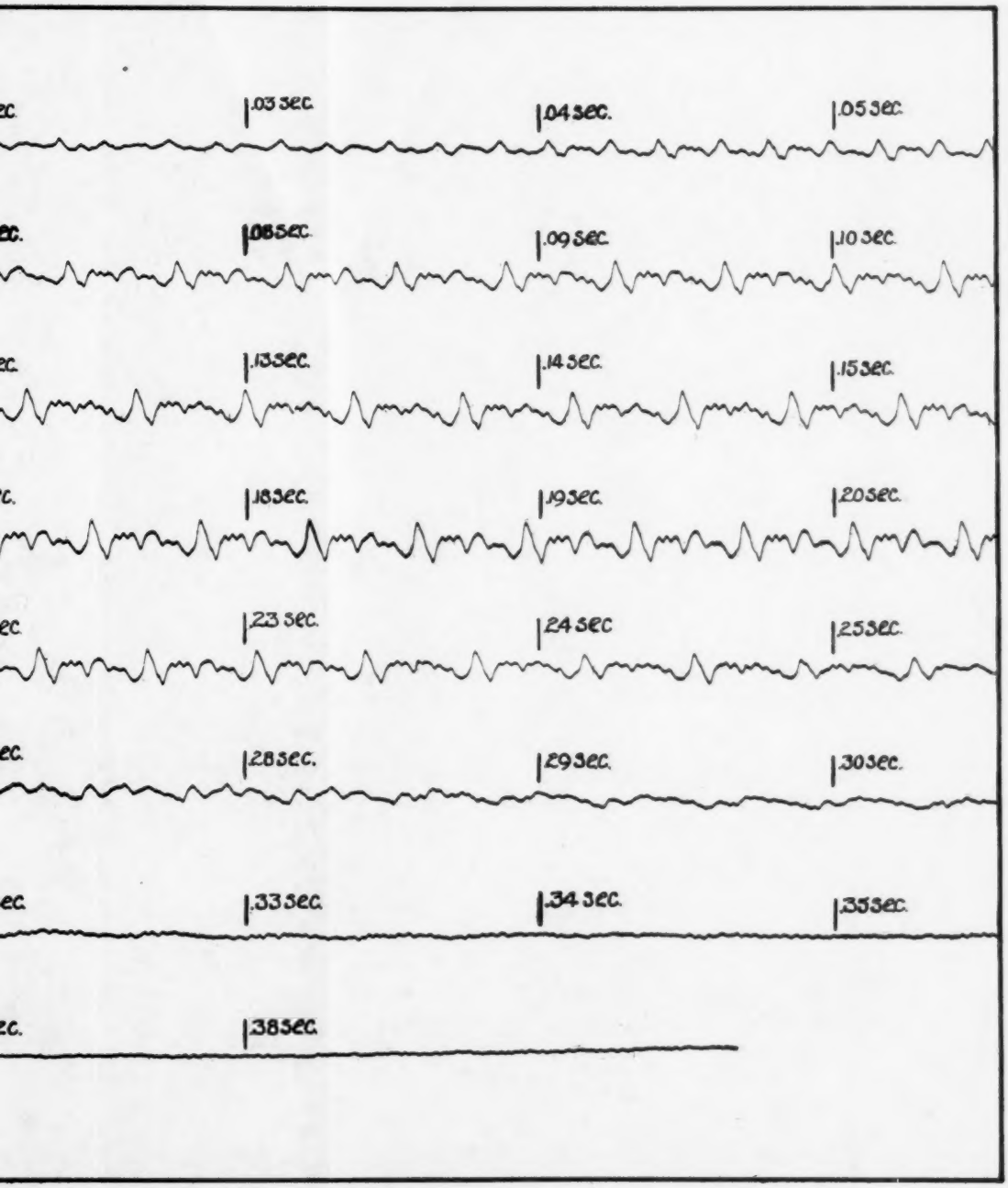


Plate No. 40—*o* as in *ton*. Sp



ton. Spoken by F.D.–Female, high-pitched

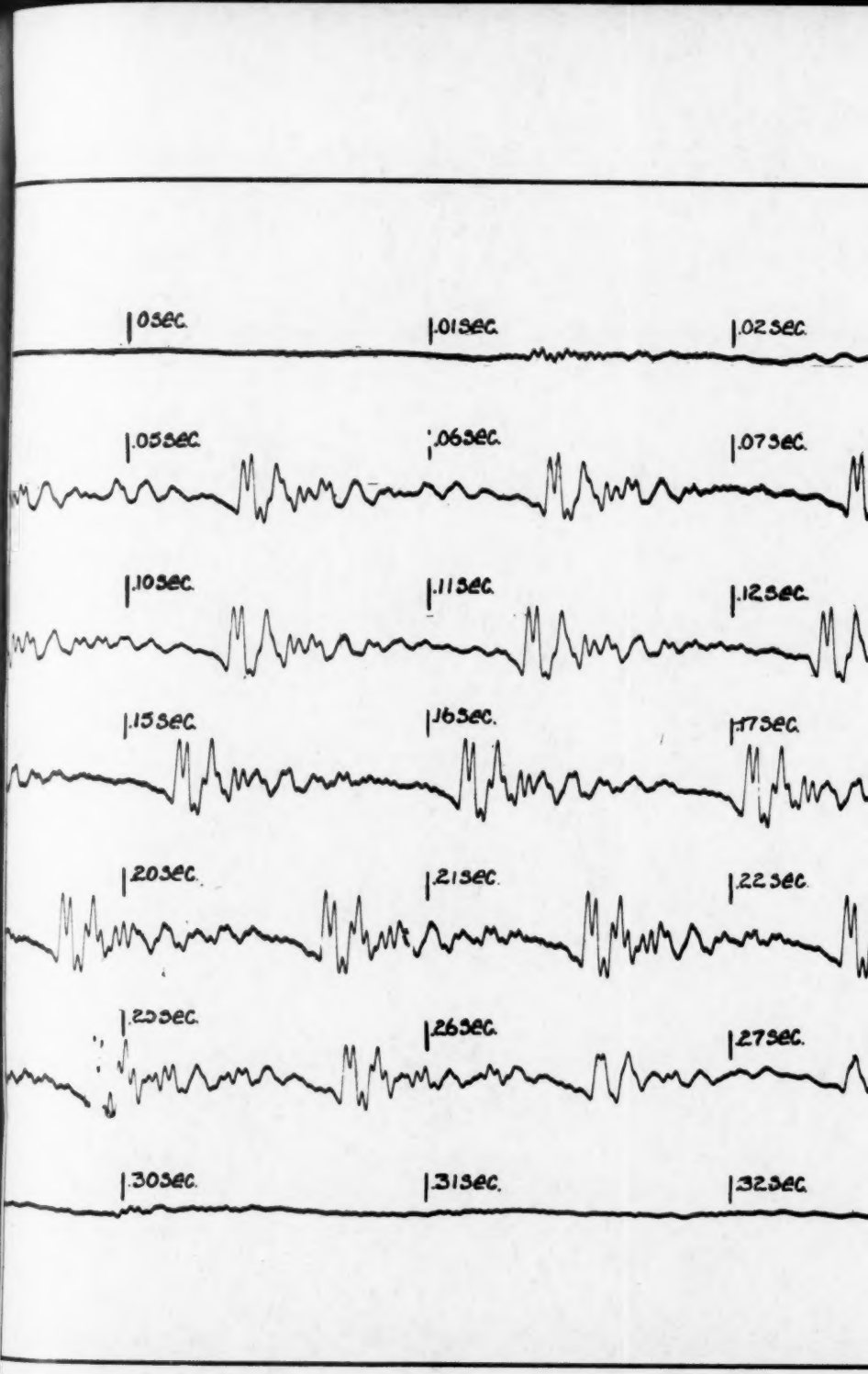
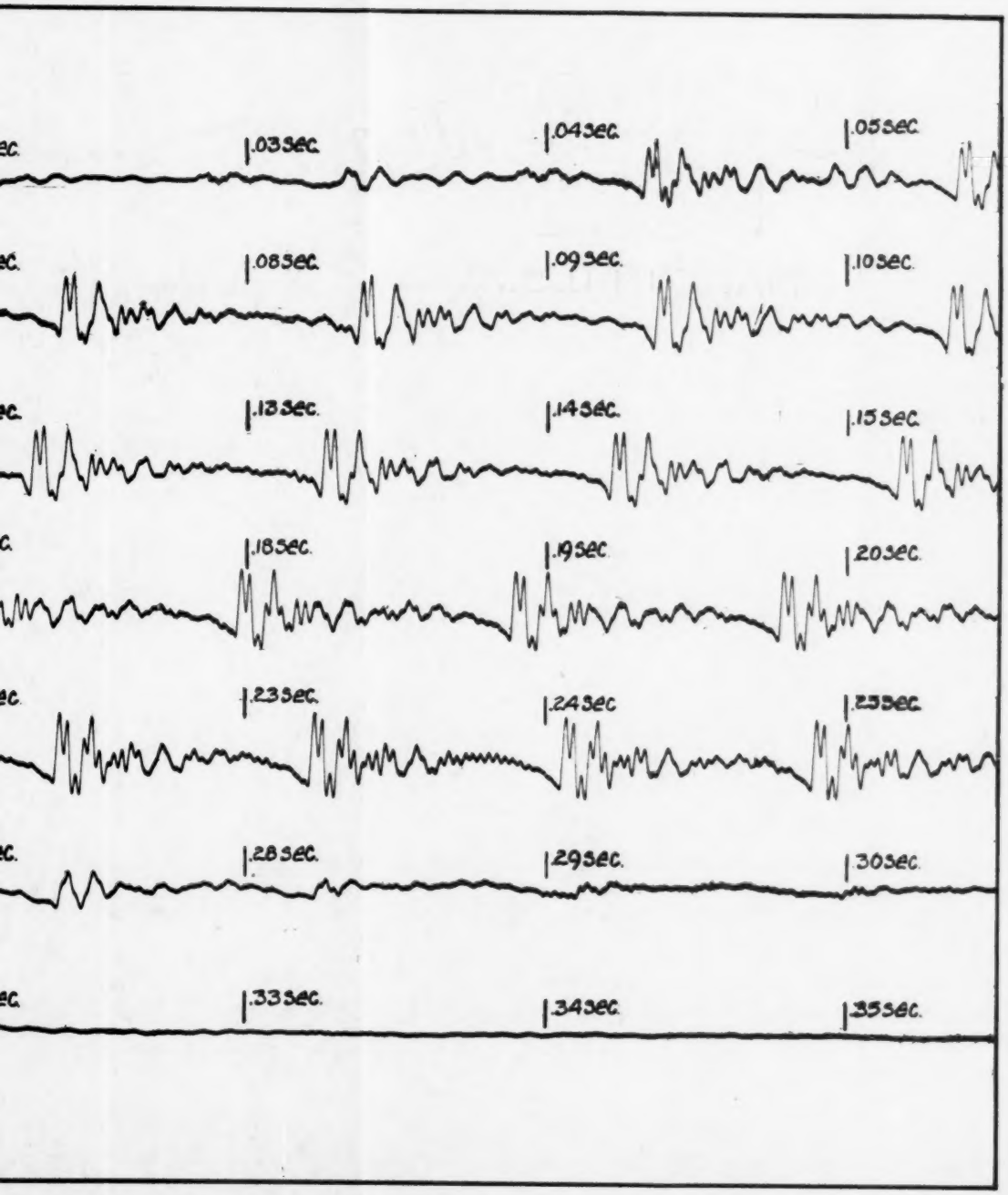


Plate No. 41—*a* as in *father*. S



ather. Spoken by M.A.–Male, low-pitched

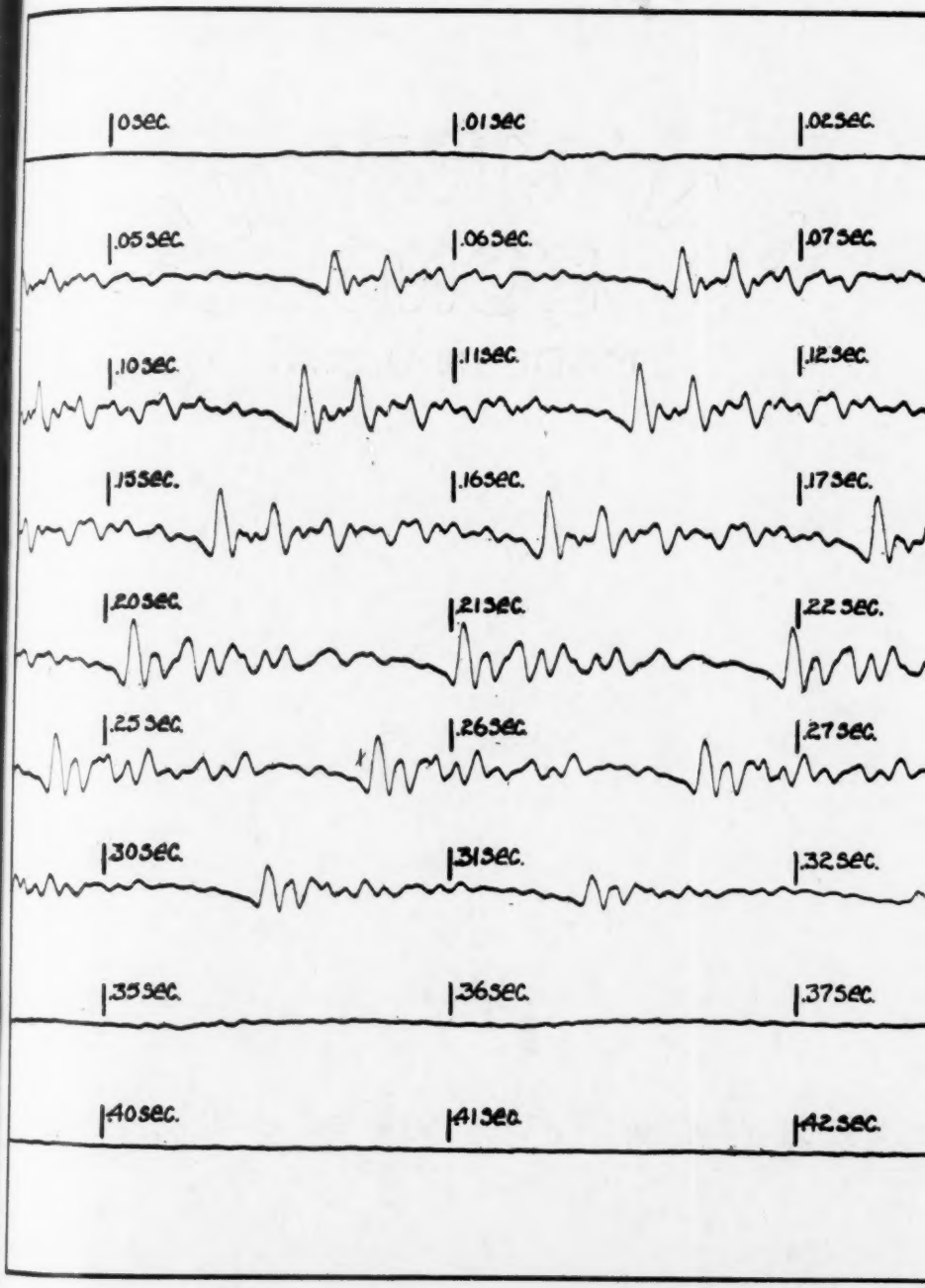
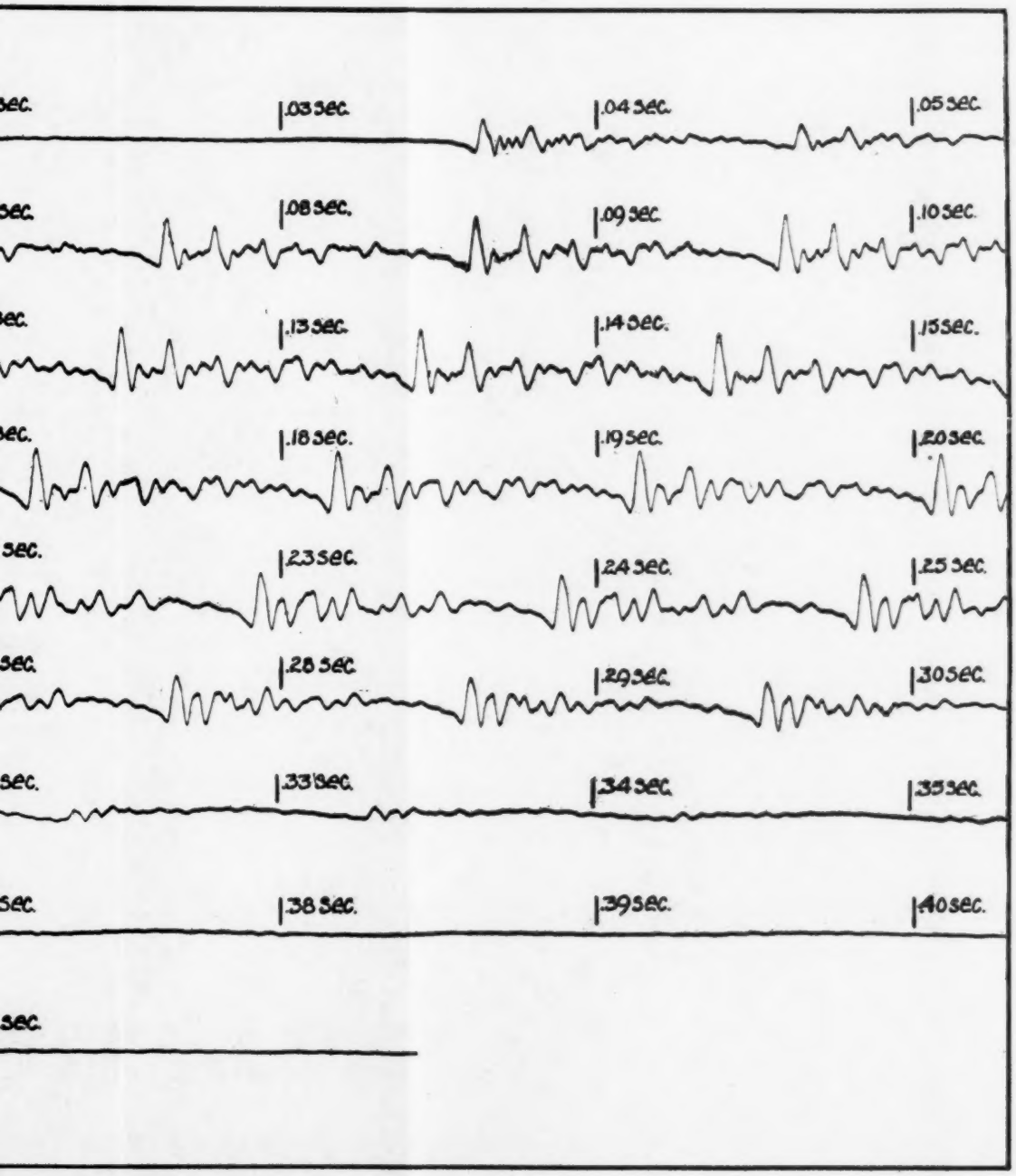


Plate No. 49—*ar* as in *part.*



in *part*. Spoken by M.A.–Male, low-pitched

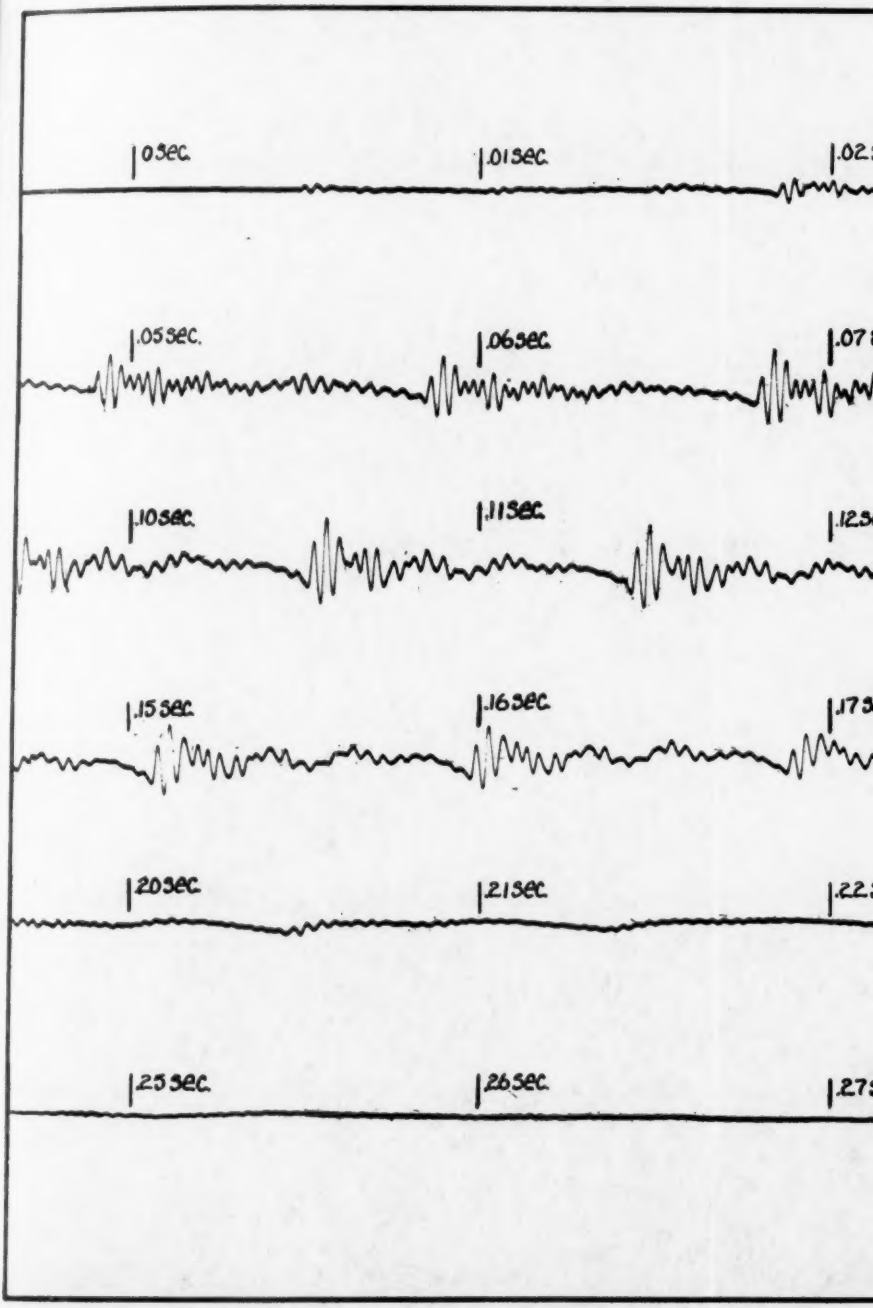
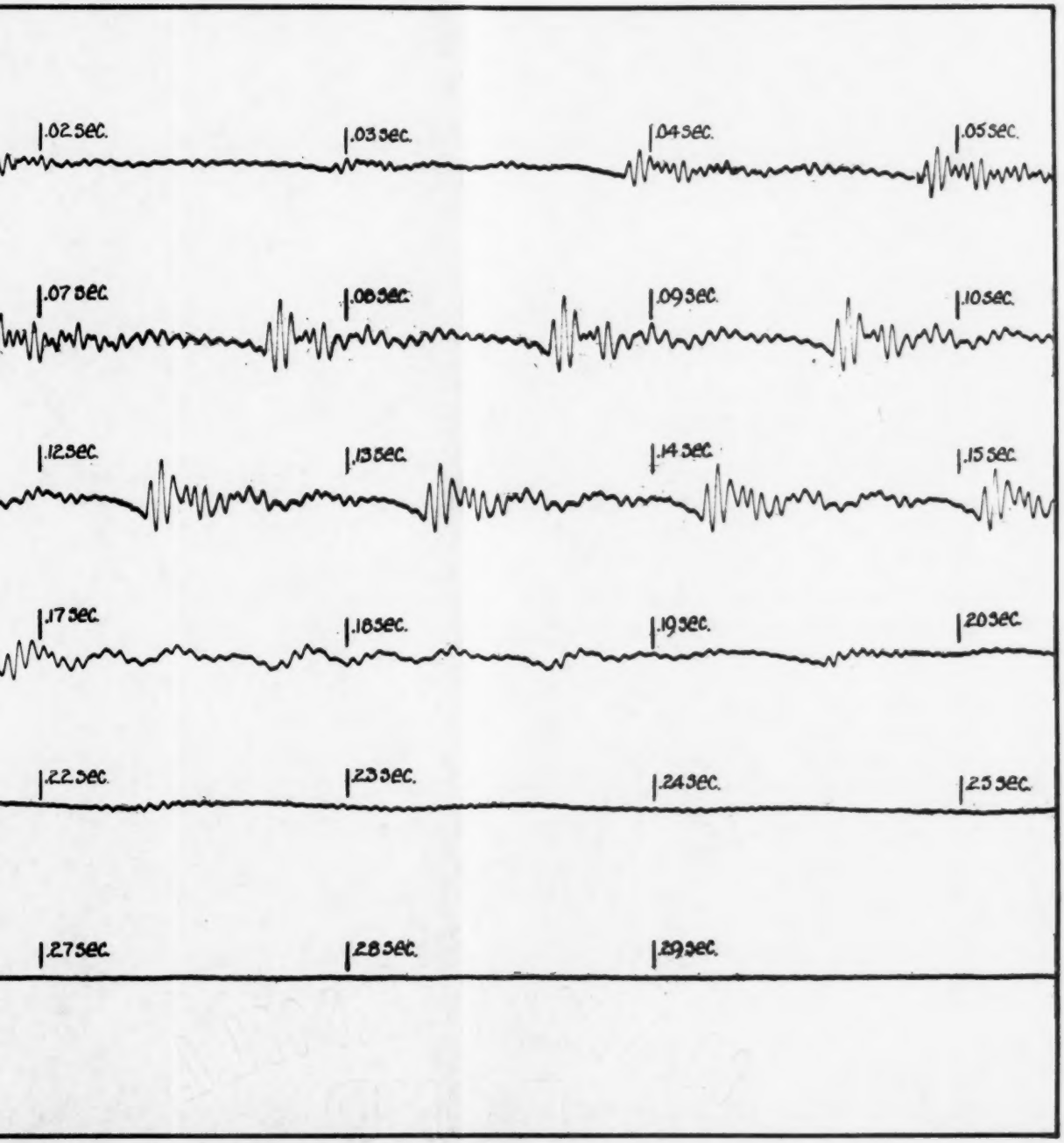


Plate No. 89—*i* a



o. 89—*i* as in *tip*. Spoken by M.A.–Male, low-pitched

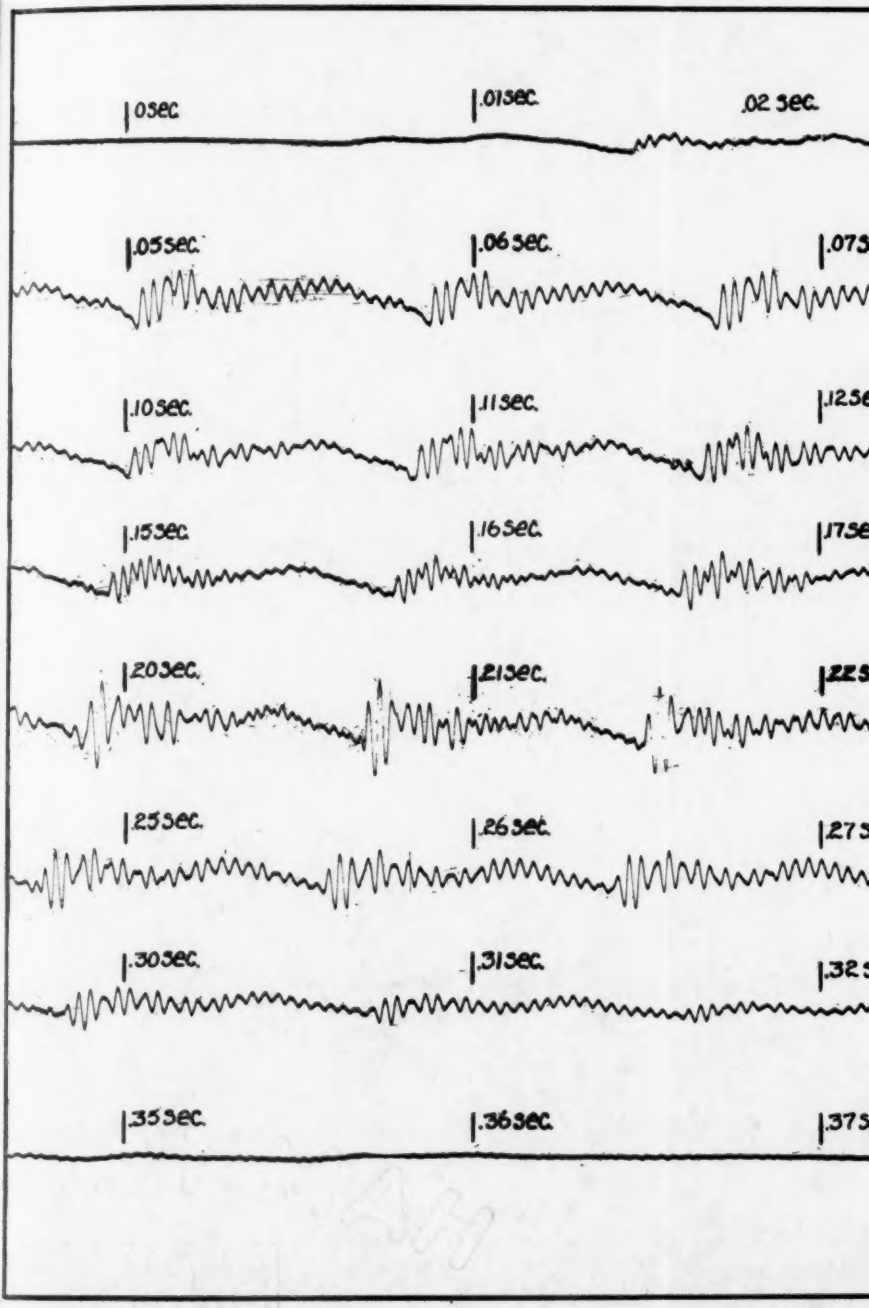

.03sec.
.013sec.
.023sec.
.053sec.
.063sec.
.073
.103sec.
.113sec.
.123
.153sec.
.163sec.
.173
.203sec.
.213sec.
.223
.253sec.
.263sec.
.273
.303sec.
.313sec.
.323
.353sec.
.363sec.
.373

Plate N

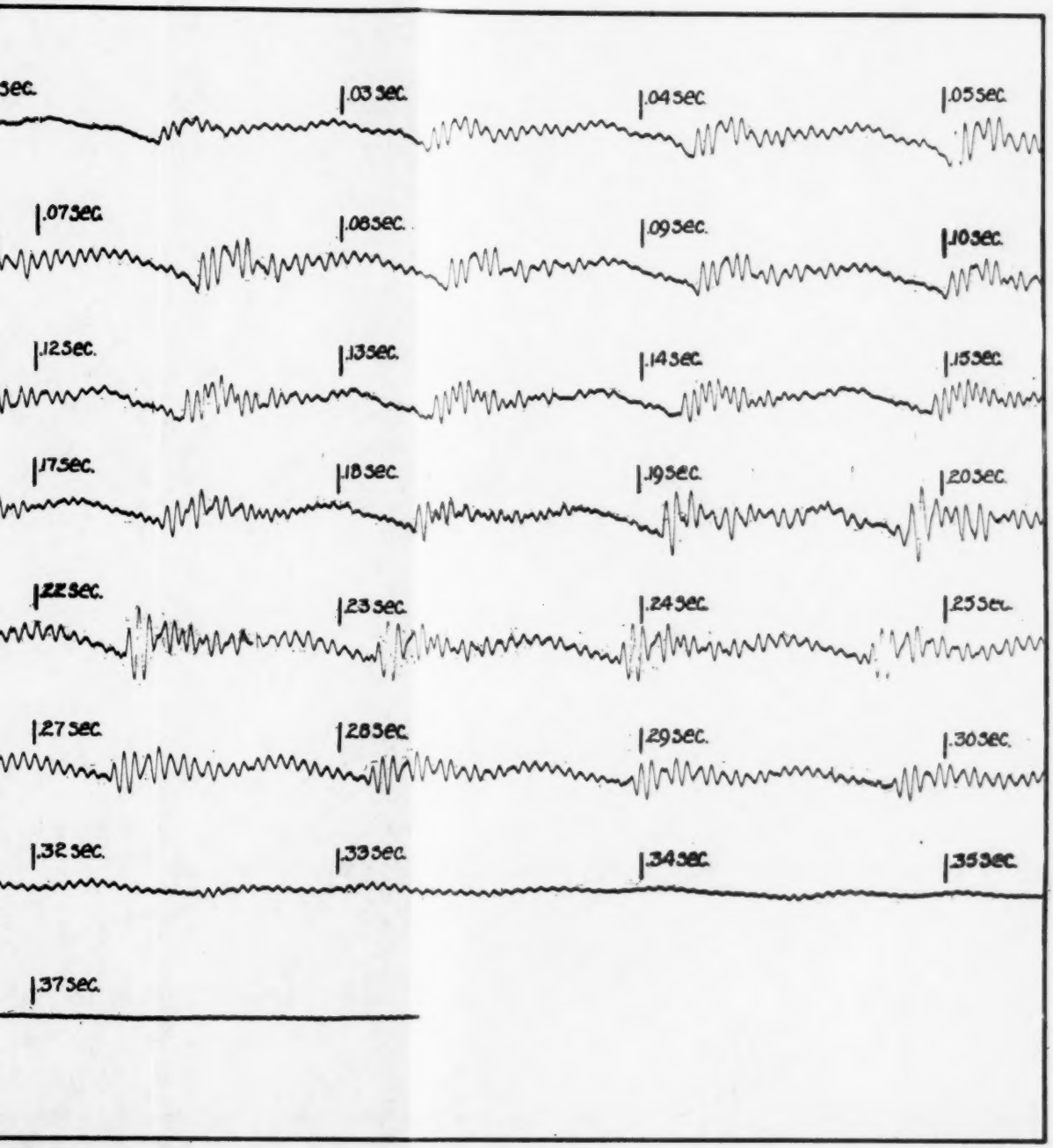


Plate No. 108—*lee*. Spoken by M.B.

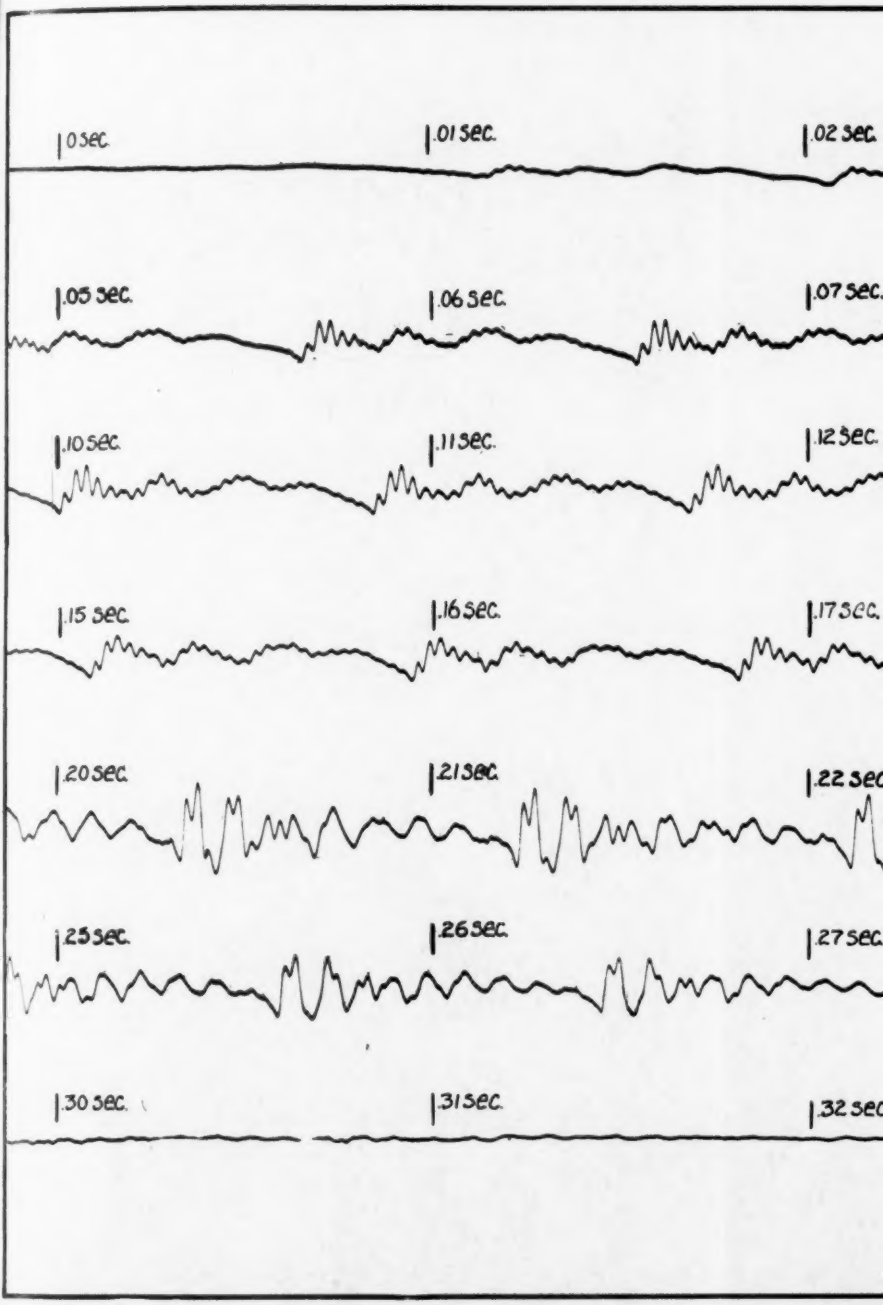

0 sec.
.01 sec.
.02 sec.
.05 sec.
.06 sec.
.07 sec.
.10 sec.
.11 sec.
.12 sec.
.15 sec.
.16 sec.
.17 sec.
.20 sec.
.21 sec.
.22 sec
.25 sec.
.26 sec.
.27 sec.
.30 sec.
.31 sec.
.32 sec

Plate N

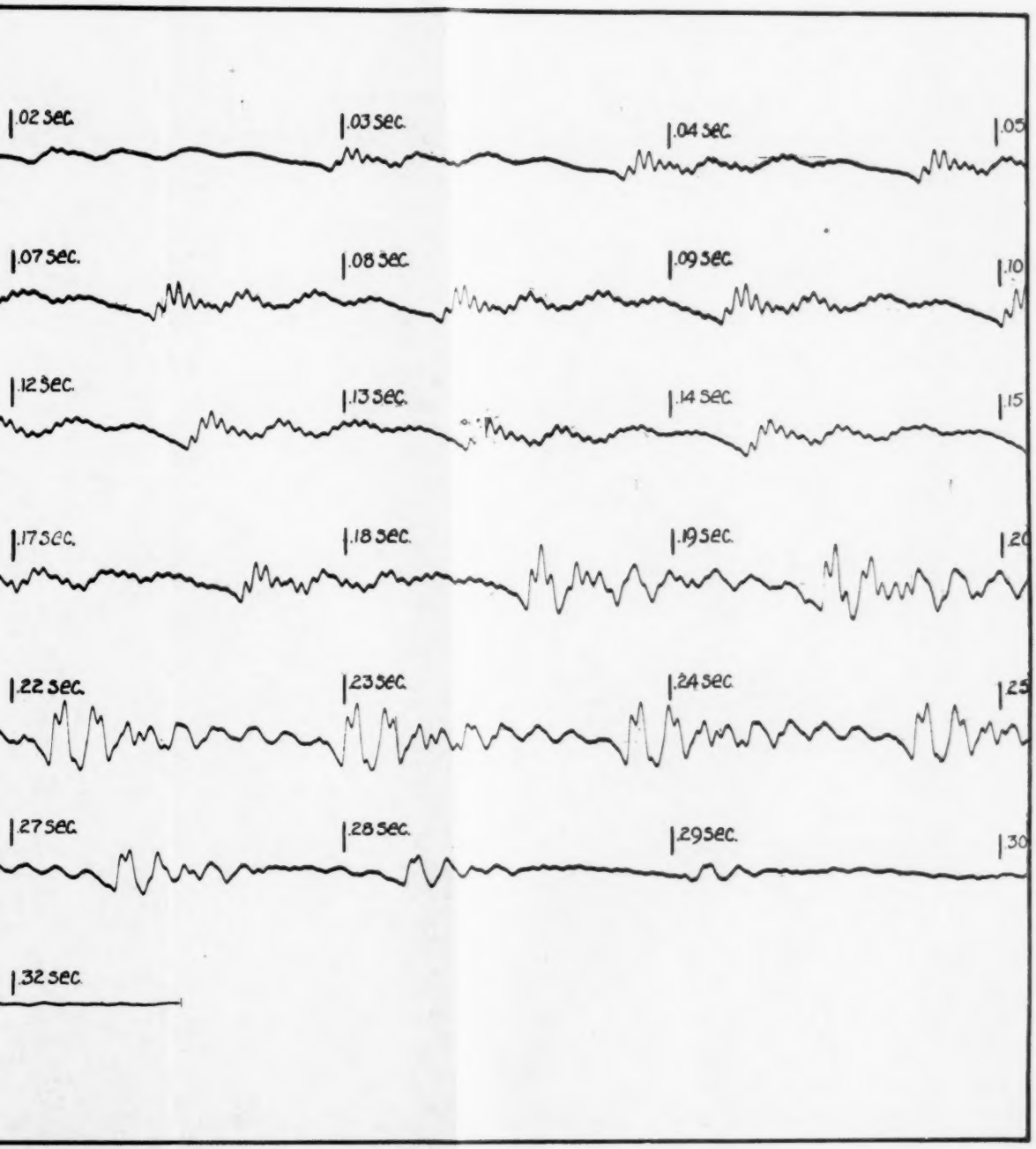


Plate No. 110—*Ia.* Spoken by M.B.

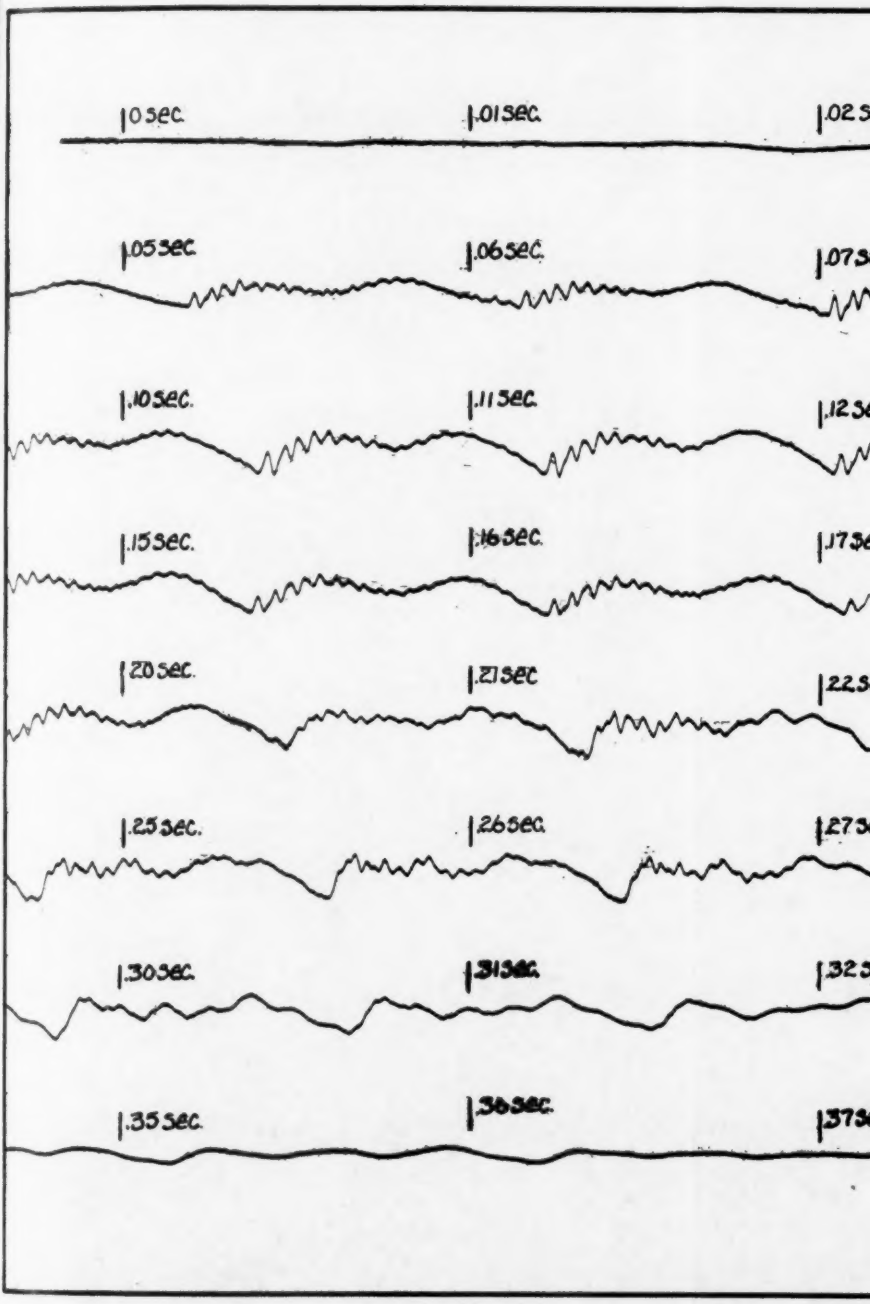

.0 sec.
.01 sec.
.05 sec.
.06 sec.
.10 sec.
.11 sec.
.15 sec.
.16 sec.
.20 sec.
.21 sec.
.25 sec.
.26 sec.
.30 sec.
.31 sec.
.35 sec.
.36 sec.

Plate N

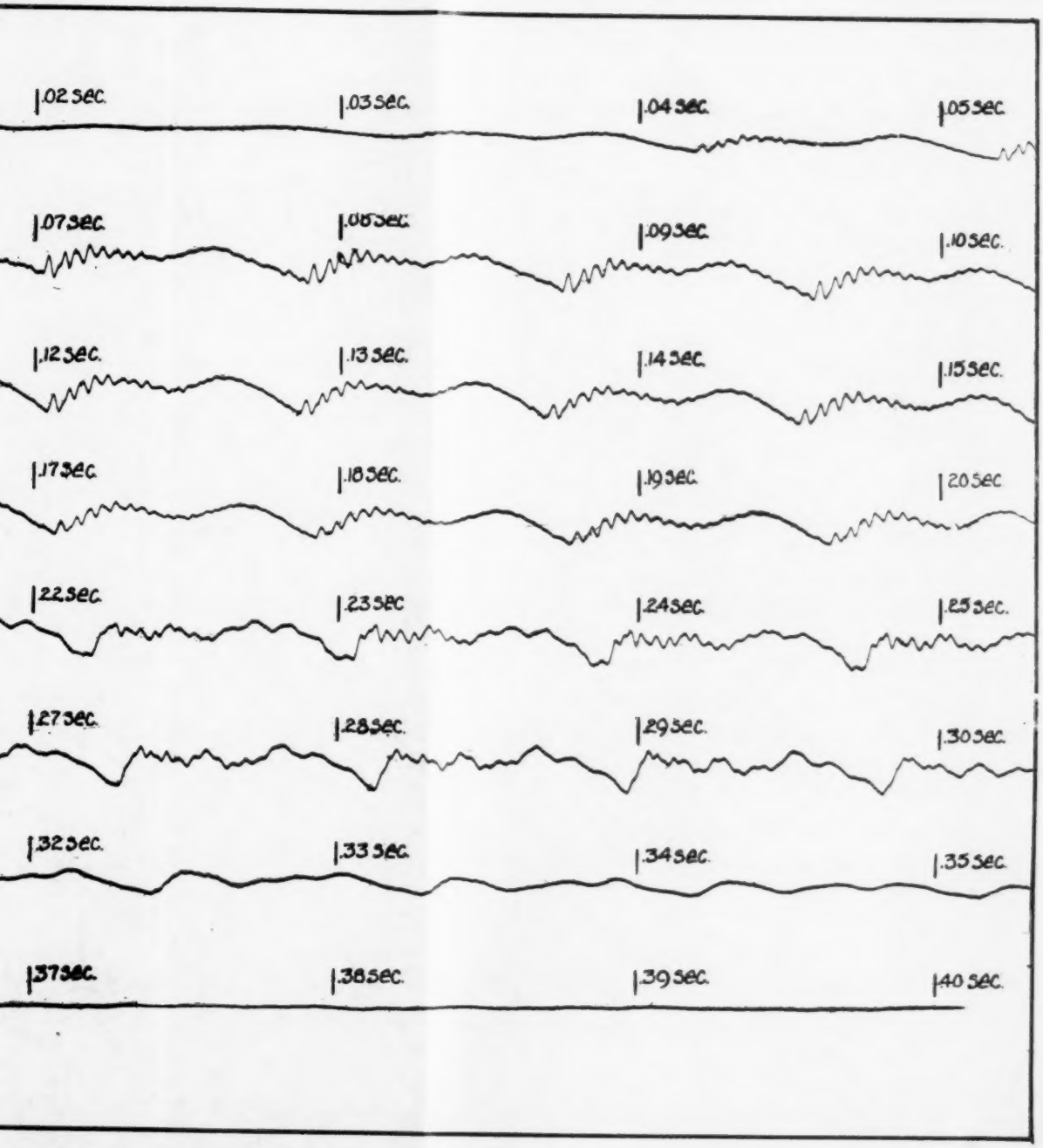


Plate No. 124—*moo*. Spoken by M.B.

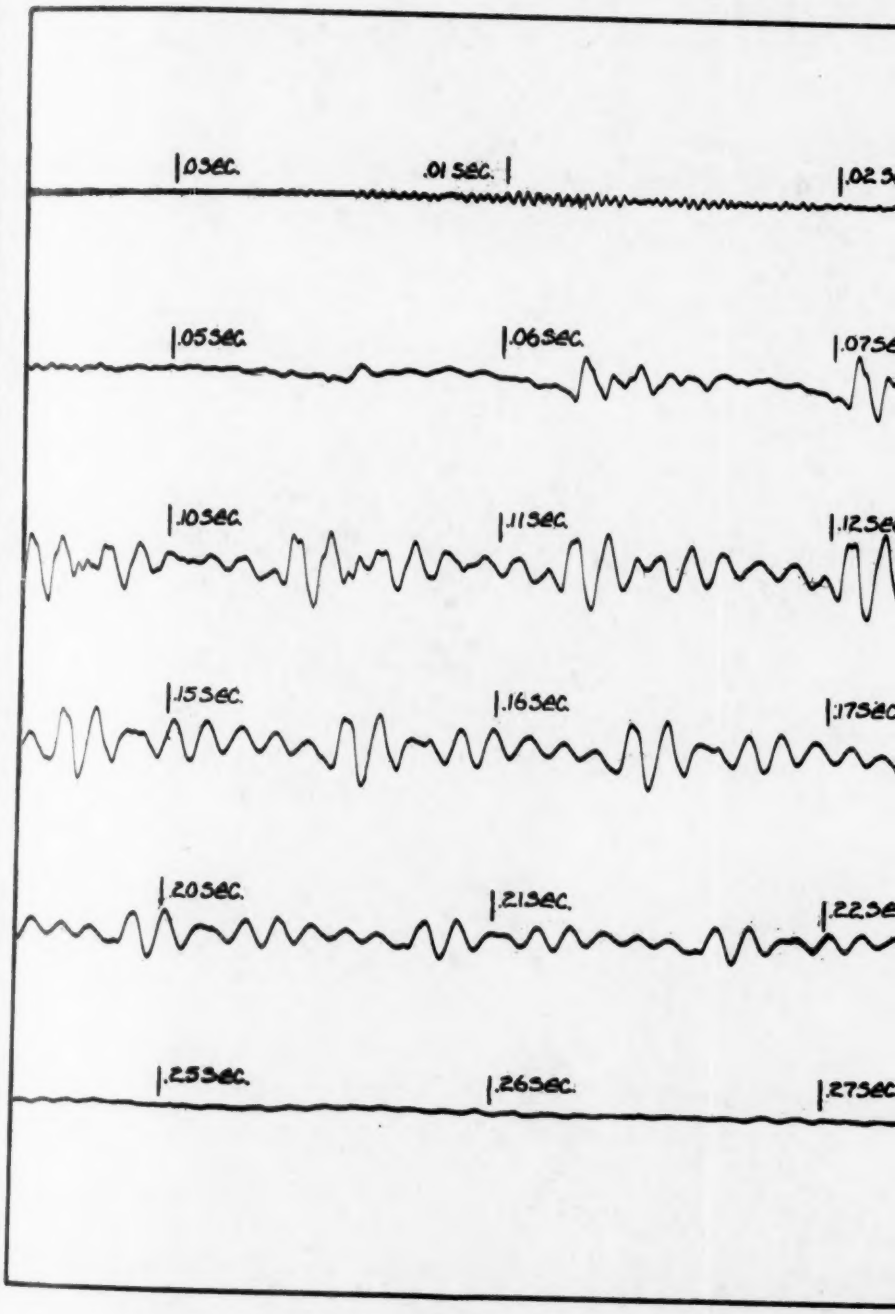
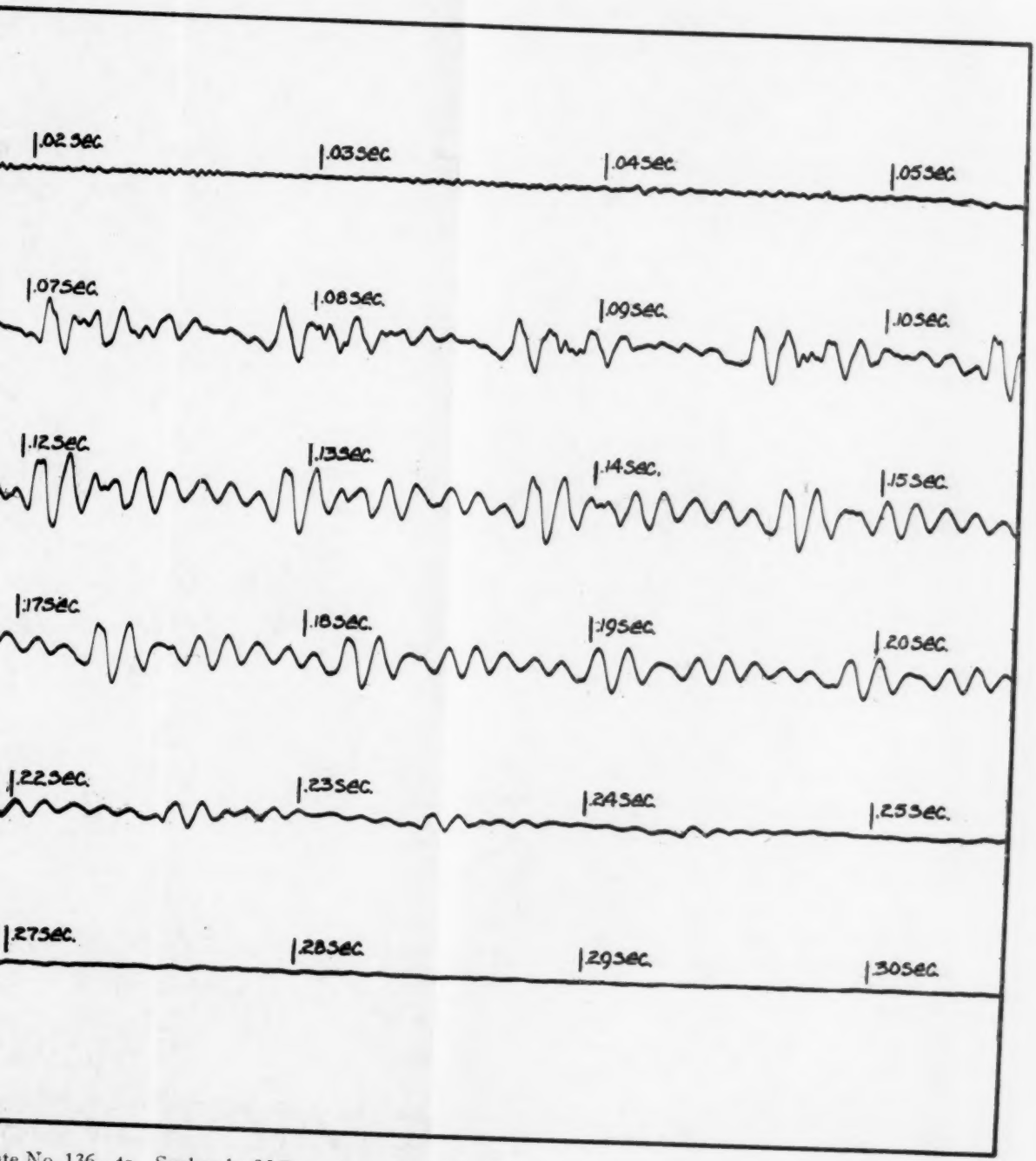

.03sec.
.01 sec.
.05sec.
.06sec.
.10sec.
.11sec.
.15sec.
.16sec.
.20sec.
.21sec.
.25sec.
.26sec.

Plate No.



te No. 136—*ta*. Spoken by M.B.

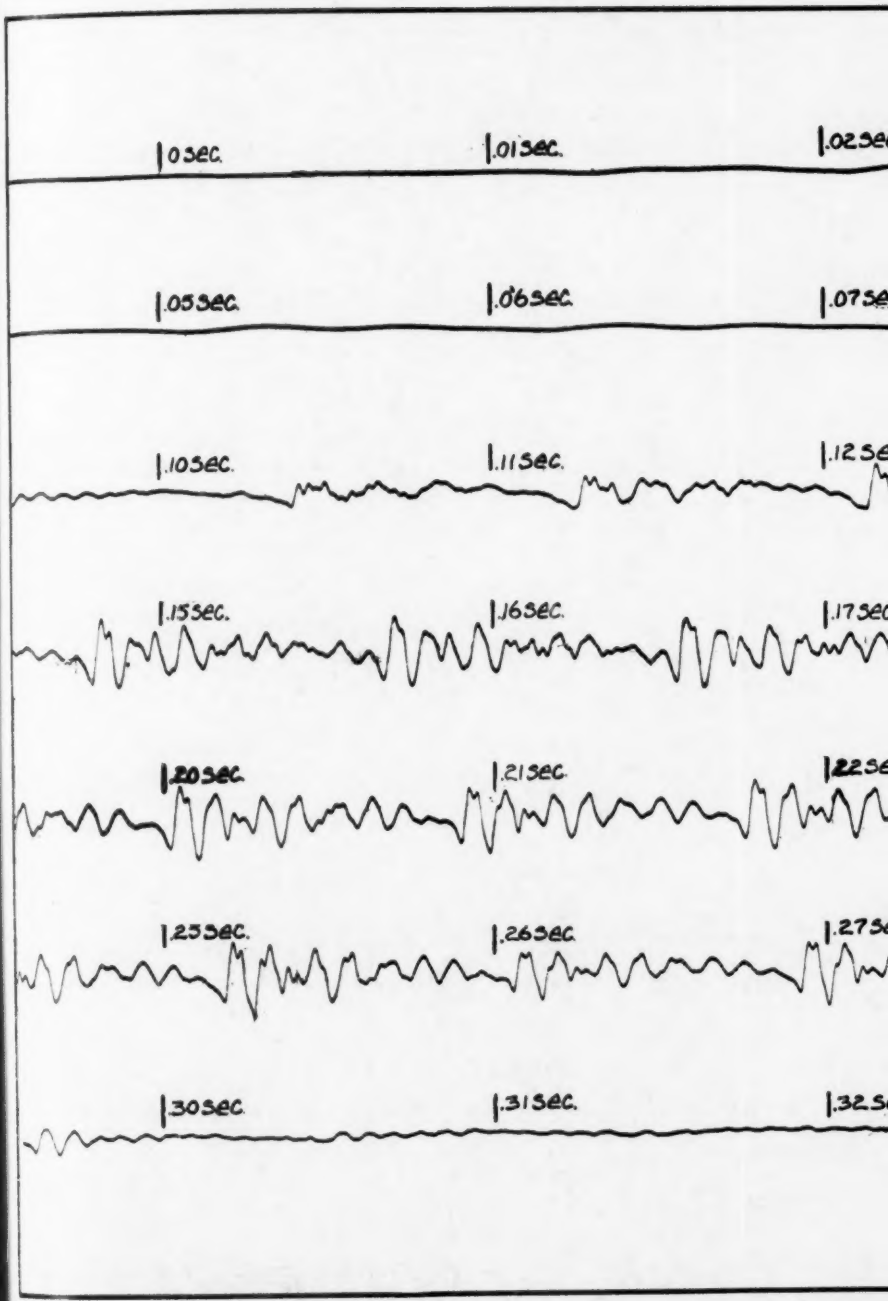
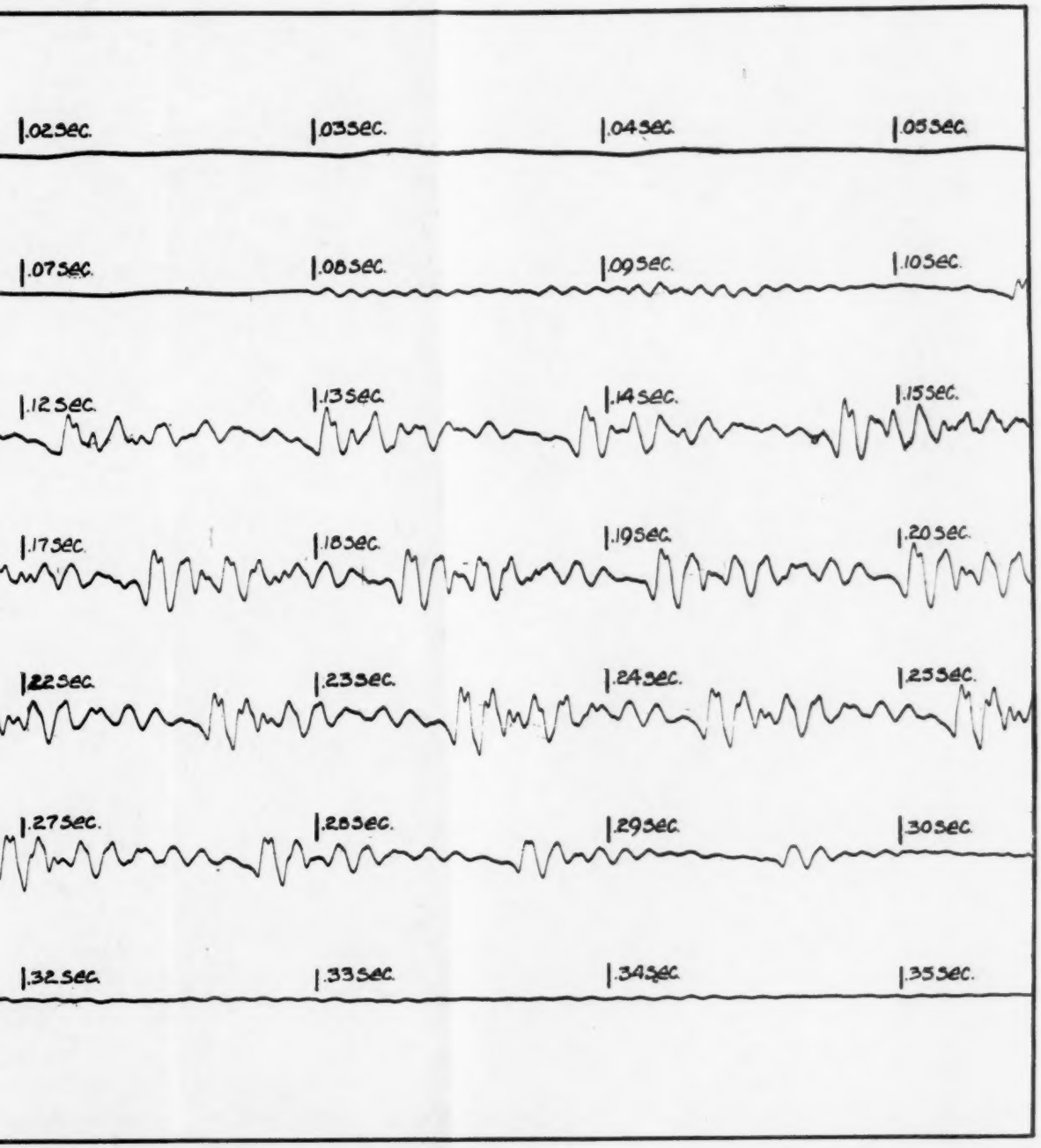
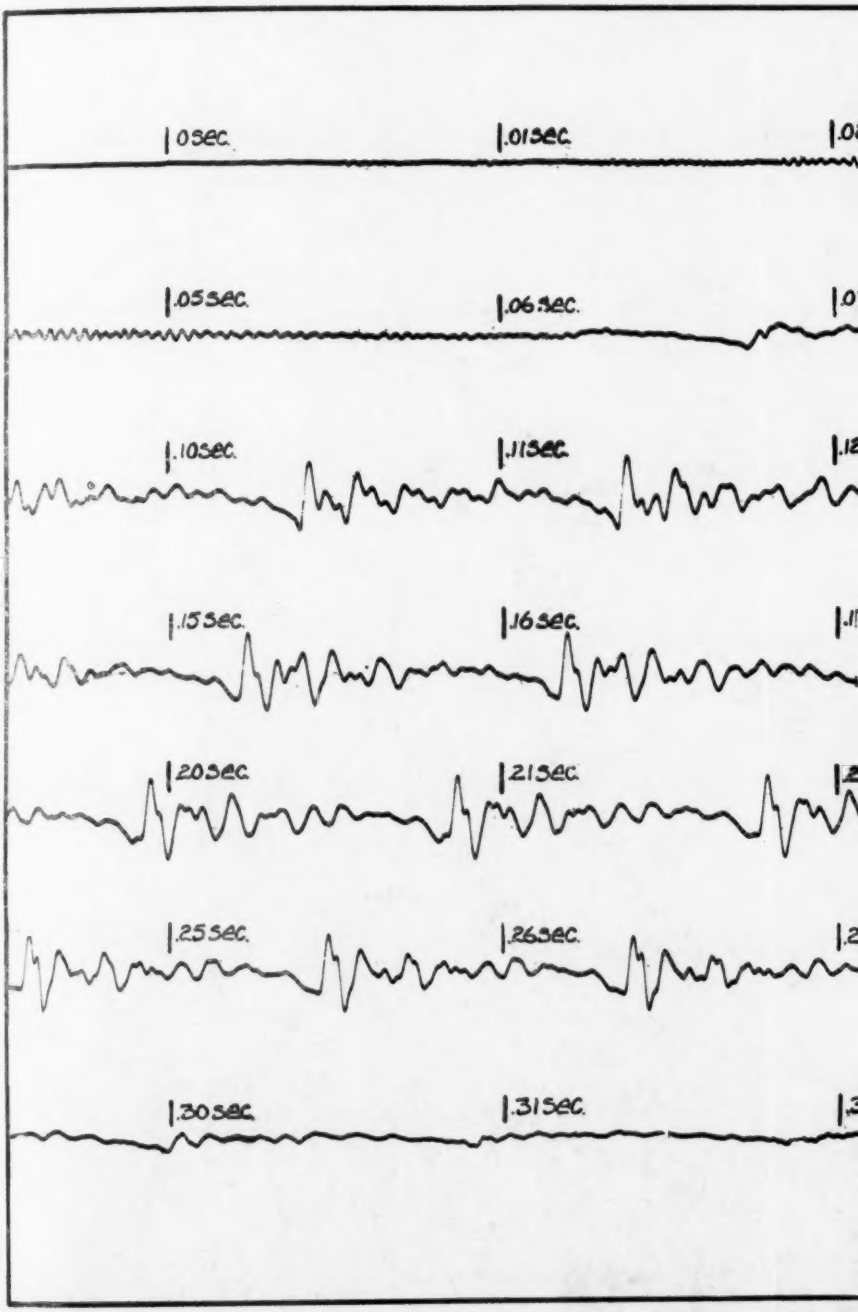

.0sec.
.01sec.
.02se
.05sec.
.06sec.
.07se
.10sec.
.11sec.
.12se
.15sec.
.16sec.
.17se
.20sec.
.21sec.
.22se
.25sec.
.26sec.
.27se
.30sec.
.31sec.
.32s

Plate No.



late No. 138—*ga*. Spoken by M.B.


.0sec.
.01sec.
.05sec.
.06sec.
.10sec.
.11sec.
.15sec.
.16sec.
.20sec.
.21sec.
.25sec.
.26sec.
.30sec.
.31sec.

Plate

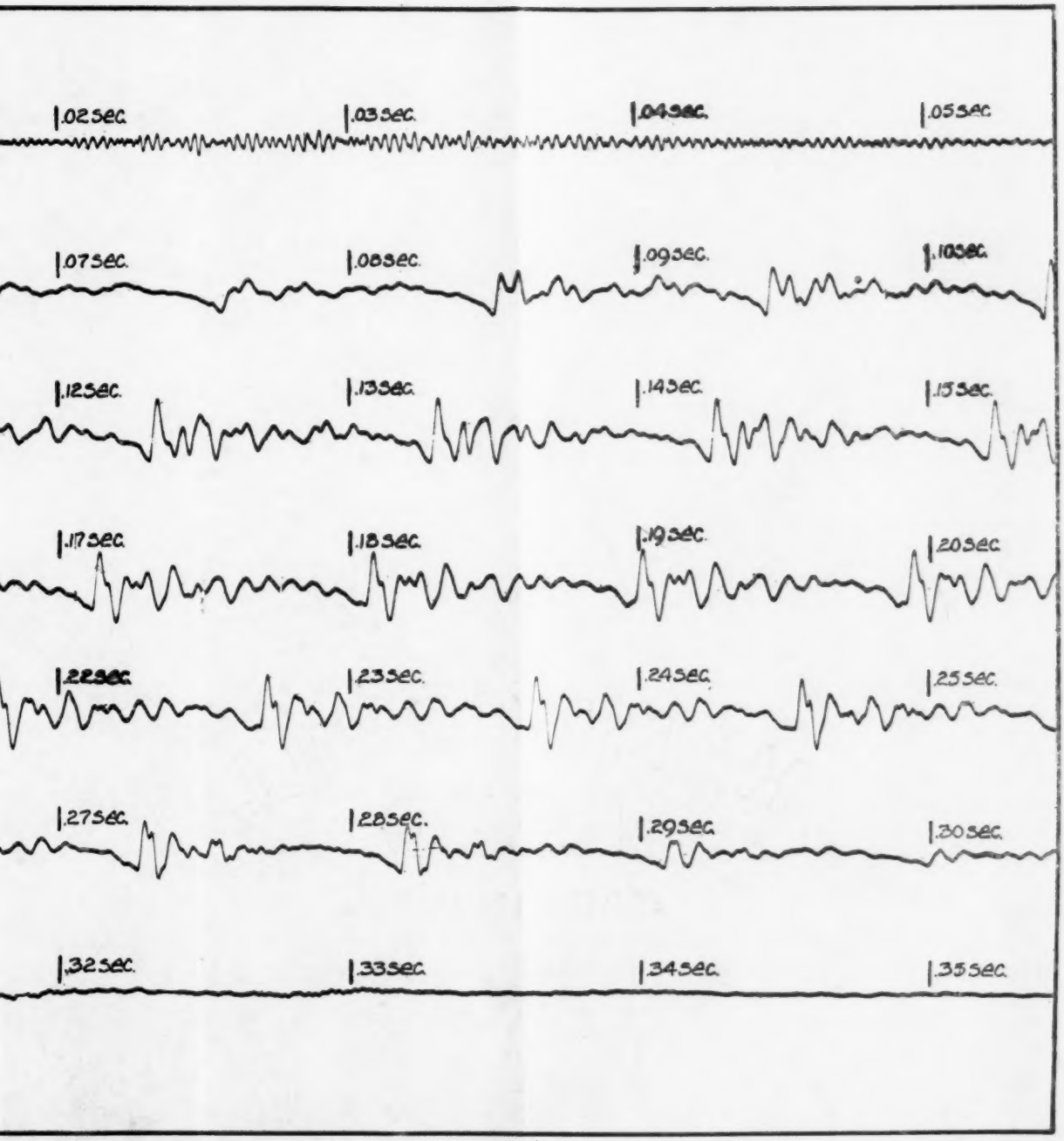
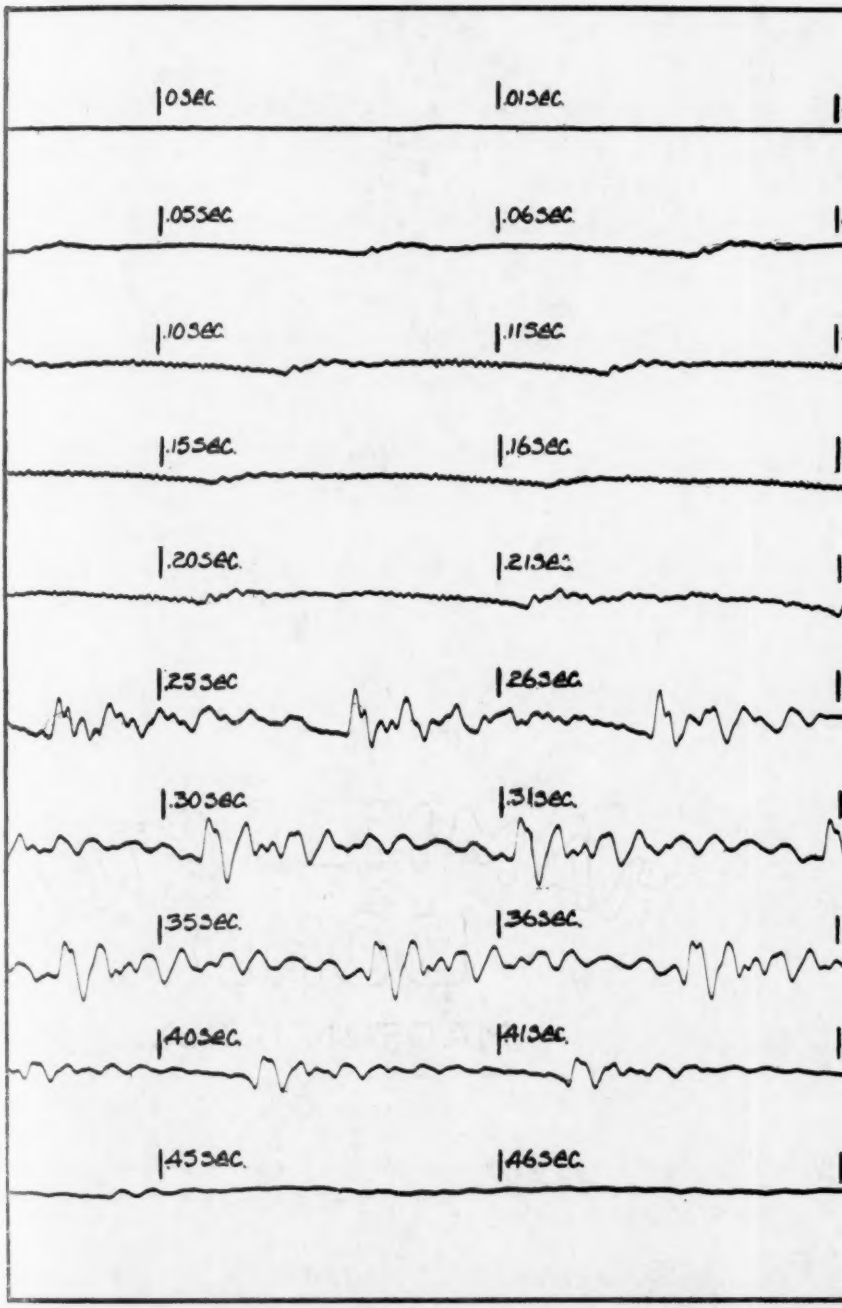


Plate No. 151—*cha*. Spoken by M.A.


.05sec.
.01sec.
.055sec.
.06sec.
.10sec.
.11sec.
.155ec.
.16sec.
.20sec.
.21sec.
.25sec
.26sec.
.30sec.
.31sec.
.35sec.
.36sec.
.40sec.
.41sec.
.45sec.
.46sec.

Plate

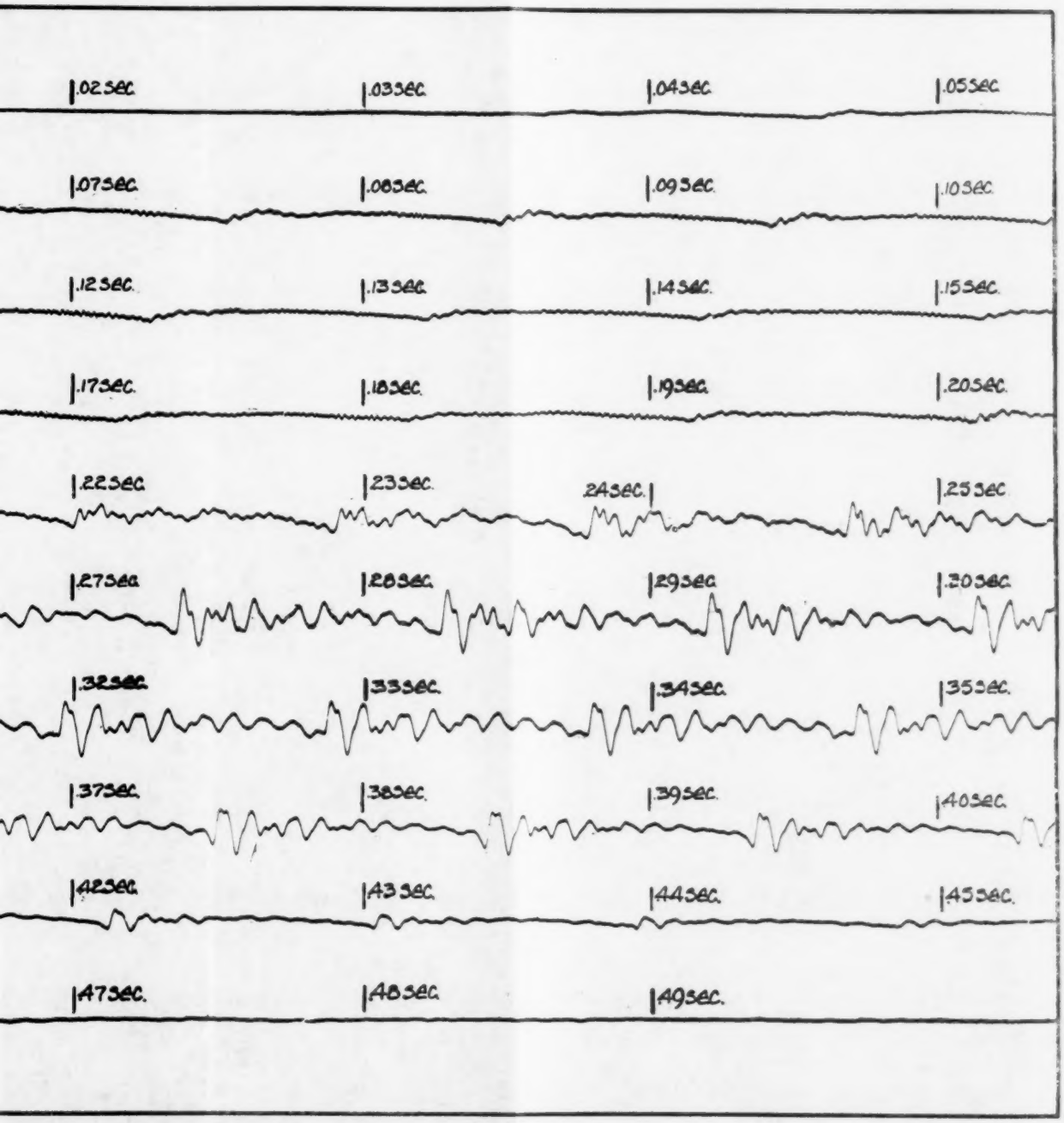


Plate No. 158—*Sa*. Spoken by M.B.

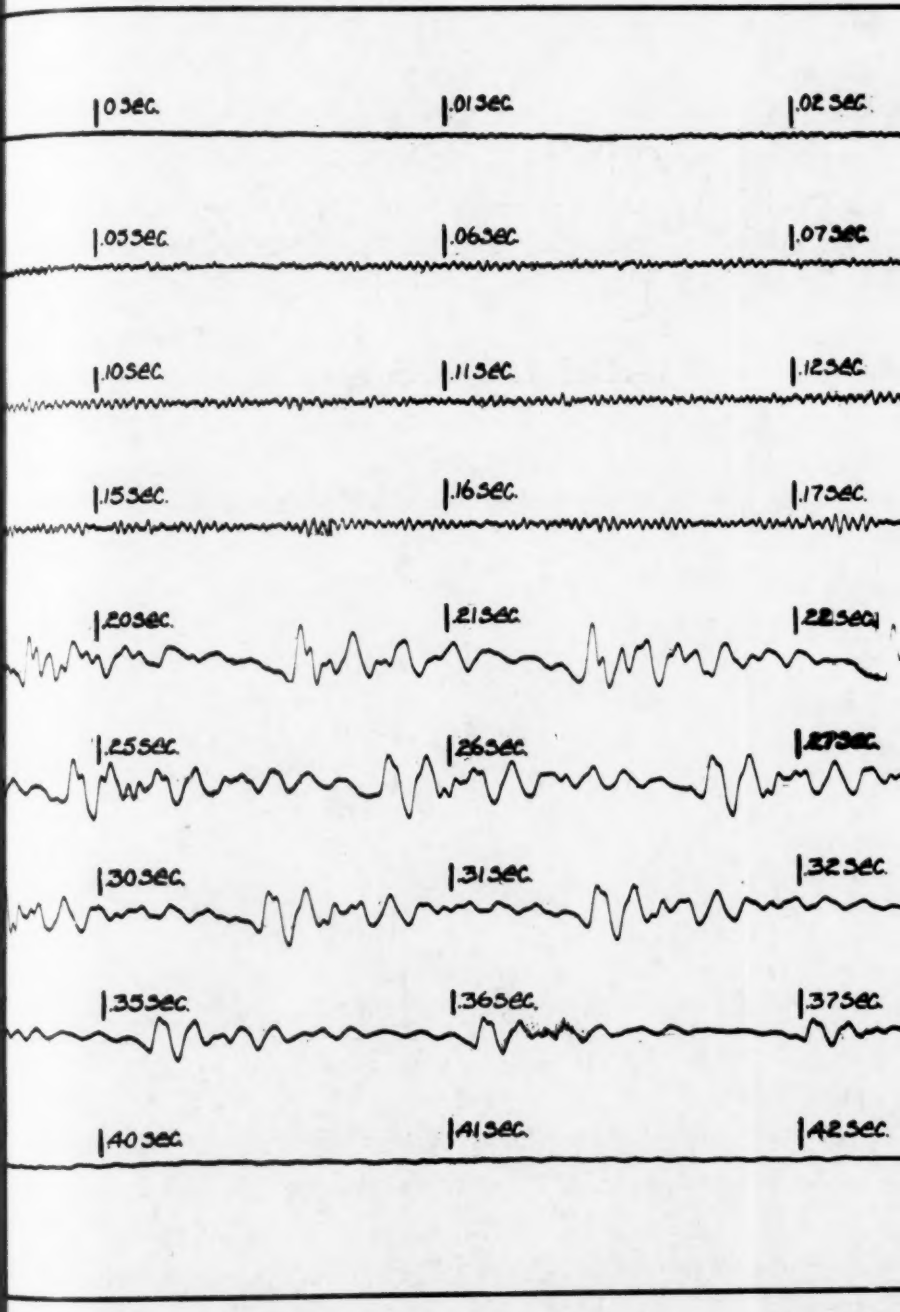
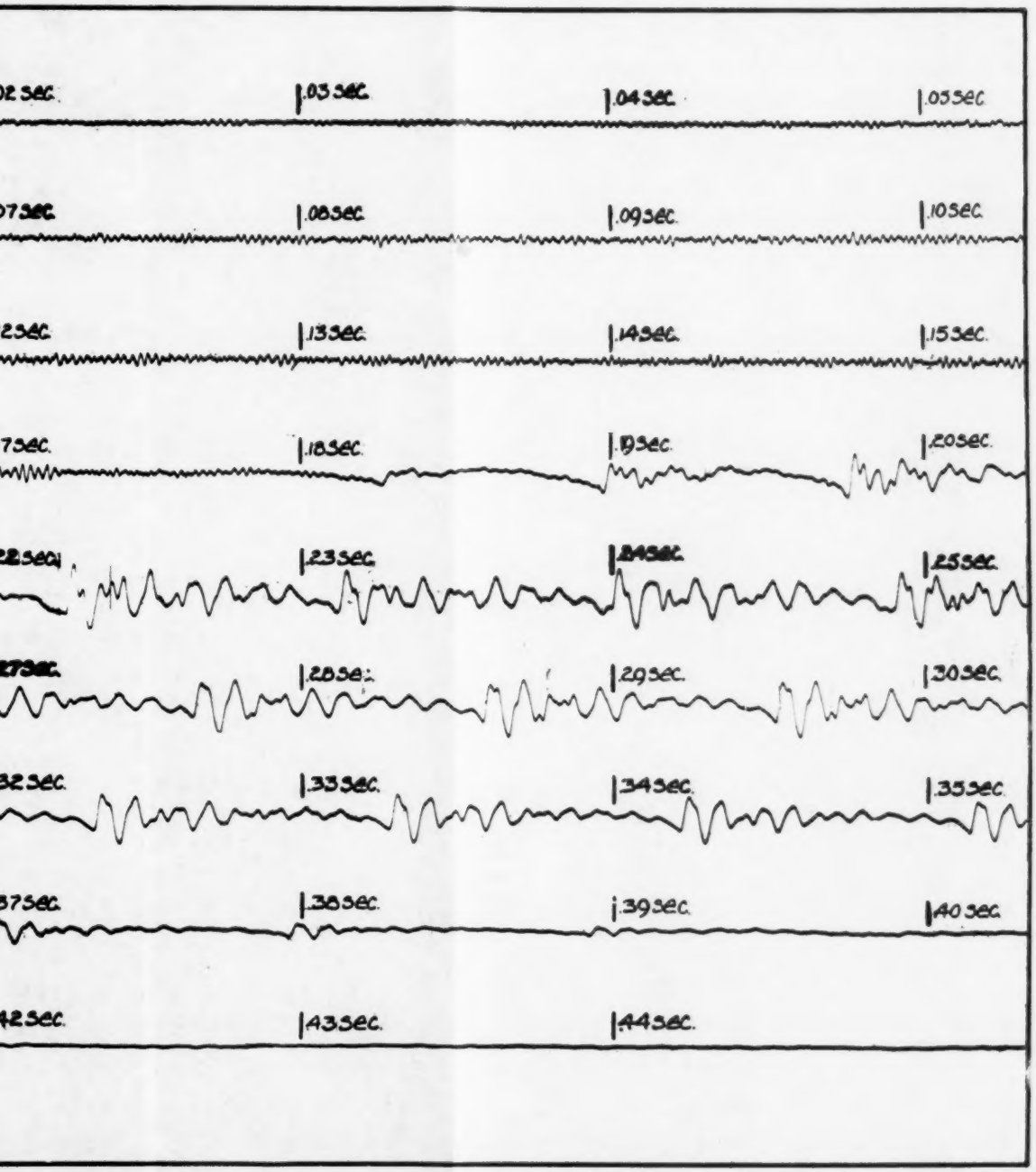

.00 sec.
.01 sec.
.02 sec.
.05 sec.
.06 sec.
.07 sec.
.10 sec.
.11 sec.
.12 sec.
.15 sec.
.16 sec.
.17 sec.
.20 sec.
.21 sec.
.22 sec.
.25 sec.
.26 sec.
.27 sec.
.30 sec.
.31 sec.
.32 sec.
.35 sec.
.36 sec.
.37 sec.
.40 sec.
.41 sec.
.42 sec.

Plate No. 16



e No. 160—*Sa*. Spoken by M.B.

with a single frequency differing from them by a small amount and detected may thereby be reduced to audible frequencies without having their interrelations of phase, amplitude or difference frequency composition, changed in any respect. For instance if the frequencies expressed above are beaten with a local constant frequency,

$$B \cos (qt+\psi)$$

the resultant lower or difference frequencies will be

$$+\frac{kBAa}{2} \cos [(p+v-q)t+\phi-\psi]$$

$$+kBA \sin [(p-q)t-\psi]$$

$$-\frac{kBAa}{2} \cos [(p-v-q)t-\phi-\psi].$$

Each one of the three components has been changed in amplitude by the same factor kB representing the efficiency of detection. Each

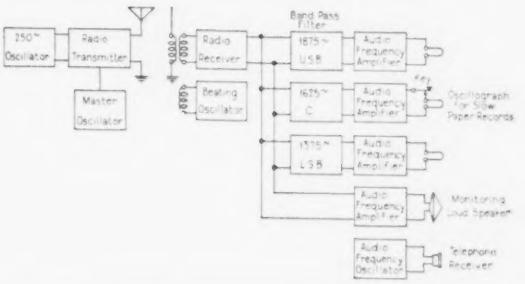


Fig. 11—Diagram of system used for three-frequency tests

one has been reduced in frequency by exactly the same amount $\frac{q}{2\pi}$ and each has had its instantaneous phase shifted by an angle $-\psi$. Relative to each other they remain unchanged.

In our actual case the carrier frequency $\frac{p}{2\pi}$ was 610 kc. The modulating frequency $\frac{v}{2\pi}$ was 250 cycles and the beating frequency $\frac{q}{2\pi}$ was 608,375 kc. so that the resulting three audio frequencies were 1,875 cycles, 1,625 cycles and 1,375 cycles.

As indicated in Fig. 11 in order to make a record of these signals they are separated at the receiver by means of band filters. These filters and others similar in type for other modulating frequencies

were designed and made by the Bell Telephone Laboratories especially for this work. The band filters used for the purpose of selecting the carrier and side-band frequencies had a cutoff of 40 Transmission Units 250 cycles from the mid-band frequency.

These cutoffs together with the position in the frequency range of the pass bands of the filters preclude any troubles from cross modulation of the radio carrier and side bands during the beating down process. The products of such cross modulation would be frequencies which are multiples of 250 cycles and these cannot pass the filters. On the other hand the beaten down frequencies will pass practically intact, since as has been shown by the previously described single frequency tests, each of the three frequencies received although subjected to amplitude modulation by fading, represents only a very narrow band of frequencies for which the filter pass bands were of adequate width.

As the modulating tone was carefully calibrated to 250 cycles and the filters adjusted to transmit the frequencies specified, it was only necessary to transmit the carrier while adjusting the receiving beating oscillator. The following procedure for this adjustment was found to be very successful. A local audio frequency oscillator was set to the reduced carrier frequency of 1,625 cycles, and its output connected to a telephone receiver. The audio beat note from the radio signal and local beating oscillator was reproduced by a loud speaker and its frequency adjusted to zero beat the 1,625-cycle tone from the telephone receiver.

When this adjustment had been completed the carrier was modulated with the 250-cycle tone, and the side-band signals automatically pass through their respective filters.

The signals from the outputs of the filters were amplified, and recorded separately by the three oscillograph elements. The sample records shown in Fig. 12 are representative.

Strips 1, 2 and 3 are taken from a long record obtained May 7, 1925, 3:22 a.m. The upper trace is a record of the upper side-band signal, the center trace the carrier, and the lower trace the lower side-band. Strips 4, 5 and 6 are from a section of a similar type of record made May 23, 1925, 1:06 a.m., where the carrier was modulated with a 500-cycle tone and different filters were used. In this record the upper trace is the lower side-band and the lower trace the upper side-band.

It will be noticed that the timing interruption appears only in the side-band signals, as the tone was interrupted before modulation took place, and that the amplitude of the carrier signal is not affected

by the interruption of the modulating tone. This makes it very easy to identify the side-band signals. These records give an excellent graphic picture of ordinary radio telephone transmission, bringing out the fact that three truly individual frequencies are transmitted to reproduce one.

In Fig. 12, strips 1, 2, and 3, the relative amplitudes of the three signals are very nearly in proportion to the relative amplitudes of

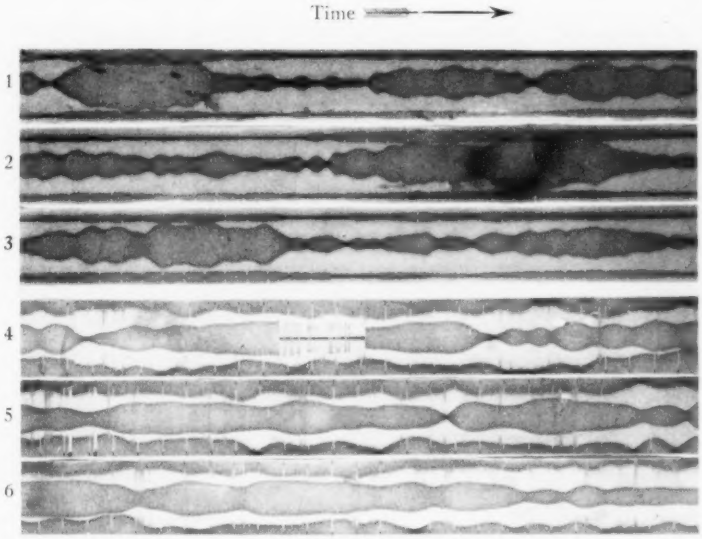


Fig. 12—Fading record showing individually the fading of carrier and side-band frequencies. Made at Riverhead, L. I. Timing interruptions in side-band signals, 5 seconds apart

the signals as they existed in the ether at the receiving point. Before this record was made a transmission characteristic of the complete receiving circuit, including the oscillograph elements, was obtained, using a local transmitter with modulated carrier for the purpose of making the measurement. The gain of the audio amplifiers at the outputs of the filters was adjusted to give substantially uniform transmission on each of the three frequencies corresponding to the carrier and side bands of the radio frequency signal.

As shown in Fig. 11, a telegraph key is placed in the circuit of the center oscillograph element, for the purpose of placing identifying signals on the records. An example of these identifying signals is

shown in Fig. 12, strip 4, which gives the date and time the record was started, July 23, 1925, 2:06 a.m. (Eastern daylight saving time).

The record in Fig. 13 is of the carrier and side-band signals with 500-cycle modulation made at Riverhead, L. I., May 25, 1925, 1:25 a.m. More gain was used in the side-band amplifiers for this record in order that the effects of fading could be brought out more prominently. In this record only half of the side-band signals were recorded, the

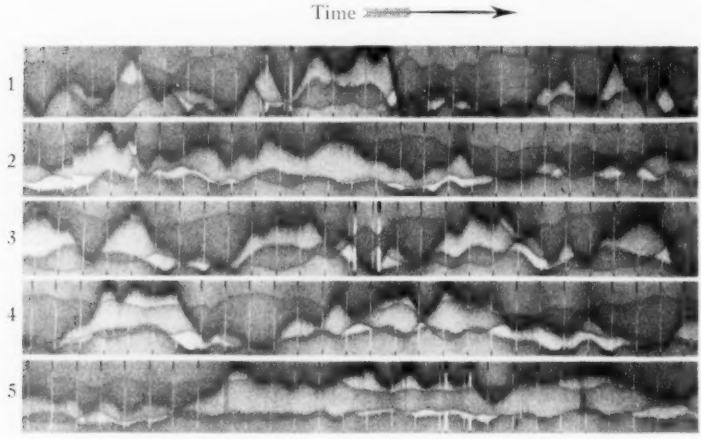


Fig. 13—Fading record of carrier and side-band signals, made at Riverhead, L. I. Timing interruptions in side-band signals, 5 seconds apart

zero reference line being at the edge of the strip. The upper trace is the upper side band, the center the carrier and lower trace the lower side band. Where the traces of the signals overlap a darker record is obtained. This record may be confusing at first but if strip 5 is examined where the amplitudes of the signals are not so large a better picture of the form of the record will be obtained.

It is obvious from these records that the carrier and side-band signals do not fade together as a unit. The carrier may pass through a zero value with still considerable amplitude in the side-band signals as in strips 1 and 3. In the first case, strip 1, the three frequencies successively fade through points of minimum signal in the order lower side-band, carrier and upper side-band; and in the second case, strip 3, the three frequencies fade through points of minimum signal in the reverse order. This is a definite indication of *selective fading*; that is, *fading is a function of frequency as well as time.*

An endeavor to form an explanation of the cause of this selective action in fading must be largely in the nature of speculation. Furthermore, since our data consist in the results of things which have happened rather than in any first hand information on the processes of the happening, the building of an explanation is a synthetic process. In general for any given set of facts it is possible to synthesize a number of explanations. Bearing this philosophy in mind we have considered various theories in connection with our observations and have concluded that simple wave interference as a major cause of the signal variations is at present the most likely explanation. While wave interference may be called a major cause it should perhaps also be called a secondary cause since the assumption of wave interference presupposes for its origin, primary causation by some physical state or configuration of the transmission medium. Speculation as to the nature of this primary cause is one stage further removed from the data contained in our oscillographic records than is the assumption of wave interference.

Since it is desirable in the remainder of this discussion to point out the evidences of wave interference, let us consider briefly the nature of this phenomenon.

To avoid any possible confusion of terms let it be said that by "wave interference" we mean a particular physical phenomenon in wave transmission and have no reference whatever to static, signals from other stations, or any other of the forms of radio noise which are commonly designated by the word "interference" when they hinder the reception of desired signals.

When two single frequency plane polarized wave trains start out at the same time from a common source and travel by different routes to meet again at a distant point the nature of disturbance at that point is determined by the relative space phases of the planes of polarization and time phases of the amplitude of the two arriving waves.

If we let E represent the vertical resultant of the electric field, which would be the only part affecting a simple vertical antenna, such as we have used in most of our tests, then

$$E = e_1 \sin 2\pi(Ft + d_1) + e_2 \sin 2\pi(Ft + d_2) \qquad (1)$$

where F is the frequency and d_1 and d_2 are the distances along the respective paths measured in wave lengths and e_1 and e_2 are the vertical components of the two waves. These two sine terms may be thought of as two vectors differing in phase.

The condition that these add giving a field

$$E=(e_1+e_2)\sin 2\pi Ft$$

$$\text{is that, } d_1-d_2=(\text{a whole number}) \tag{2}$$

that is, the difference in length of the two paths must be an exact whole number of wave lengths. The condition that the two waves cancel each other giving a field

$$E=(e_1-e_2)\sin 2\pi Ft$$

$$\text{is that, } d_1-d_2=(\text{a whole number})+\tfrac{1}{2} \tag{3}$$

that is, the difference in length of path must be an exact odd number of half wave lengths.

Thus if the two components e_1 and e_2 are equal, the resultant vertical field E will go through values ranging from (e_1+e_2) down to zero as the path lengths change relative to each other. If the two waves do not have exactly the same amplitude, the minimum value of E will be something more than zero.

Differences in attenuation of the two waves and differences in their direction of arrival will modify the relative amplitudes of e_1 and e_2 but will not modify the time relations required for minima of the resultant field E unless we assume that at the time of a minimum neither wave has an appreciable vertical component. Since the consequences of such an assumption do not accord with our experimental data we have considered that it may be left out of account in the present discussion.

This is obviously a picture which fits in very well with the simple single frequency fading records. The major maxima and minima occur when the conditions of equations (2) and (3) are met and e_1 and e_2 are nearly equal. On the other hand it seems doubtful that the picture can be so simple. If we suppose two wave paths why not three or more? Additional paths would add irregularities to the fading and it would not be necessary to assume as great a degree of irregularity in the changes in any one path. But with an increasing number of paths the various arriving waves would tend to average to a more or less constant mean value and large departures from this mean would become rare. The fact that the fading signal continually covers a large range of amplitude, with the maximum many times the minimum, definitely points toward there being but a very small number of major paths, probably not more than two.

Considering now the question of selective fading in relation to wave interference we refer back to equation (2).

If we assume the distances to be measured in any desired units and call them d_1' and d_2' our equation will still hold provided we divide each distance by the wave length measured in the same units, thus

$$\frac{d_1'-d_2'}{\lambda}=\text{a whole number}=x;$$

rearranging this and writing $\frac{V}{F}$ for λ where V equals the velocity of the waves, we have

$$F=x\left(\frac{V}{d_1'-d_2'}\right). \tag{4}$$

If now we assume $(d_1'-d_2')$ to be fixed we find that F can have a series of values which are integral multiples of $\frac{V}{d_1'-d_2'}$ which we may call the frequency spacing interval. That is, with changing frequency E will go through maximum values with frequency at a series of frequencies beginning theoretically with zero and extending upward in regular spacing to infinity.

The spacing interval is obviously that number of cycles which corresponds to the lowest finite frequency in the series, namely, the frequency for which the distance $(d_1'-d_2')$ is just one wave length since when $x=$unity equation (4) becomes

$$F_1=\frac{V}{d_1'-d_2'}=\text{the spacing interval.} \tag{5}$$

By using the same process on equation (3) we find that E has minimum or zero values at another series of frequencies having the same spacing interval but lying midway between the frequencies at which maxima occur.

Thus it is apparent that with fixed path length difference the amplitude of the field E will be different for different frequencies, ranging from maxima of (e_1+e_2) down to minima of zero if the polarization planes and amplitudes of the two vertical components are equal.

Furthermore, still thinking of equation (1) as representing two vectors, it is evident that the phase of the resultant field is different for different frequencies even though these different frequencies had exactly the same starting phase at the source.

If the paths are changing with time, the field at a given point, as has already been pointed out, will go through time fluctuations. Another way to look at this is that there is a space pattern of maxima

and minima and as the paths change the plane section of the pattern taken by the surface of the earth wanders so that at any one point the field is continually fading in and out as the maxima and minima glide by it. Each frequency has its own pattern differing from those of its neighboring frequencies in such a way that at any given point the relation between amplitude and frequency is that just discussed

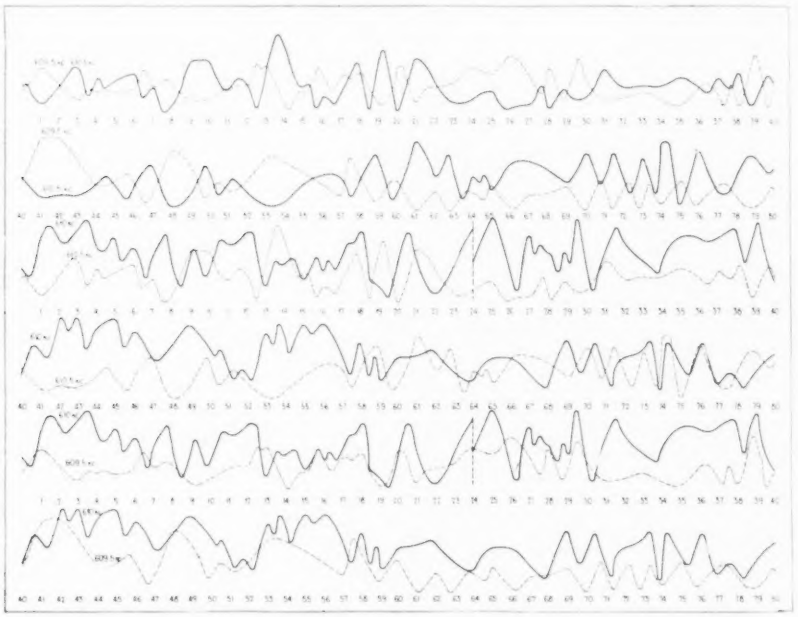


Fig. 14—Plotted curves of signal amplitudes condensing a long fading record, part of which is shown in Fig. 13. Numbers along time axis correspond to successive 25 second timing interruptions

above. Thus as the paths change and the patterns shift the different frequencies fade not simultaneously but progressively.

In the above analysis of wave interference it has been assumed that all frequencies traveled from transmitter to receiver over a given path in the same elapsed time. This does not mean that they necessarily follow exactly the same route on this path since they might follow somewhat different routes of equal length or if their transmission velocities were different they might follow different routes of unequal length and still come within the definition of a "path." It seems reasonable to assume that over the width of an

ordinary transmitted band the various frequencies are treated alike by the medium and the simple assumption that they follow the same route with the same velocity is justified. If none of these assumptions is correct but the departure is not large the effect will be merely to introduce slight irregularities into the spacing interval and the general nature of the result will not be changed.

Let us now examine more closely the record, a part of which is shown in Fig. 13. A portion of this has been condensed into the curves of Fig. 14. One unit along the time axes of these curves represents a 25-second interval.

To obtain these curves the amplitude of the signal has been scaled off and plotted, ignoring all the minor irregularities. From this record the relative fading characteristics of these single frequency signals 500 cycles apart are more easily seen, and it is possible to contrast the time of occurrence of points of minimum signal for any pair of them.

For the frequency difference of 500 cycles (610.5–610 and 610–609.5) these times are obviously quite different but there is no clearly

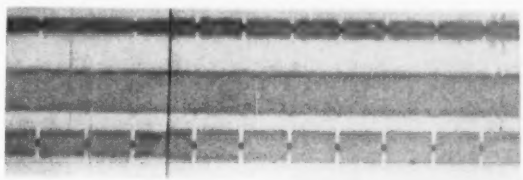


Fig. 15—Daytime record of carrier and side-band signals

discernible relation between them. The curves for 1000-cycle difference (609.5–610.5), however, show a striking relation in that the maxima and minima of the two are opposed fairly regularly over the entire 33-minute interval covered by the plot. This means that when one frequency has a minimum amplitude the other has a maximum and vice versa. Certainly this suggests a wave interference involving only two major paths whose difference in length is such that the spacing interval is 2,000 cycles. The path difference appears to be changing somewhat irregularly but at an average rate of the order of one wave length (or approximately 500 meters) per minute.

Before speculating further on the numerical values which may be derived from this part of the data we had perhaps best consider some other records of a somewhat different kind which are better adapted to provide such values. But first let us reiterate that these are *night-time* effects.

During the day signals substantially uniform in amplitude are received. An example of the type of transmission obtained in the daytime is given in Fig. 15, which is a record of the carrier and side-band signals received with substantially the same terminal conditions with the exception of the time as that existed when the records shown in Fig. 12 were made.

The abrupt change in the amplitude of the side-band signals was due to an intentional change at the transmitter in the input level of the tone modulating the carrier, and accordingly the amplitude of the carrier did not change. The timing interval is 5 seconds.

Band Fading Records

The familiar fading record is limited to two axes, amplitude and time. So far we have extended this cramped perspective somewhat by observing as many as three separate fading records spaced at audio-frequency intervals along the frequency axis. Even these three narrow lookouts upon the wide range of ether transmission have indicated amplitude relations along the frequency axis which

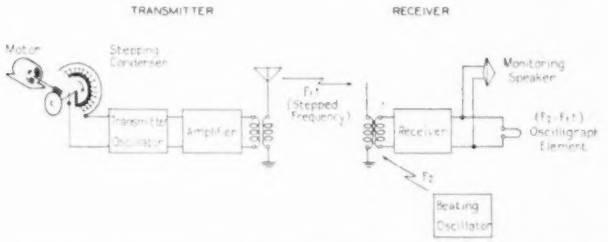


Fig. 16—Diagram of system used to obtain records of selective fading or "band fading" records

promise to open a new line of attack upon the problem of night-time fading. But the desirability of knowing what takes place in the interval unrevealed by these cracks in the fence becomes obvious. We should like to know the relative amplitude of frequencies over a wide band, and the change in this relation with time.

Since it is not a simple matter to record simultaneously the amplitude of a large number of waves of frequencies separated by say one hundred cycles in the radio-frequency range a single frequency in combination with a frequency stepping device at the transmitter has been adopted. The circuit arrangement is shown diagrammatically in Fig. 16. The rotary contactor bringing into the circuit suc-

cessively a total of fifteen small condensers across the main condenser of the transmitter oscillator shifts the frequency in steps over an adjustable range. The contactor is rotated at the rate of nine revolutions a minute, which is sufficiently slow to show definite steps in the oscillograph record. At the receiving end a local oscillator supplies a radio-frequency wave for beating the incoming frequencies down to values within the audible range.

A long oscillograph record of this stepped frequency gives a sort of moving picture of the fading for the entire band covered. A sample of such a record is shown in Fig. 17 with alternate pictures in the series removed to simplify the relations, since by reason of

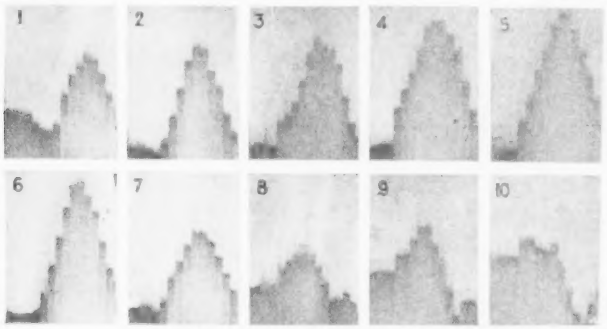


Fig. 17—Sample band fading record

the two-way traversal of the frequency band successive pictures are reversed. If a series of such built-up pictures as these could be taken rapidly on moving picture film, and projected successively upon a screen we should have before us an animated view of band fading. And according to the results of experimental investigation the subject offers a lively theme for such a presentation. The peaks and depressions glide nervously back and forth across the setting. The successive pictures of Fig. 17 (which, by the way, were selected for their half-tone reproduction possibilities rather than as first class examples of the records taken) illustrate a rather leisurely movement of this sort. These ten built-up photographs cover a period of slightly more than one minute. In the first seven pictures a depression appears at the left, while in the last three this depression seems to have made an exit followed by the simultaneous entrance of another from the opposite wing of the stage. Evidence of such

organized spacing of the minima is present in all of these night-time band fading records. As has already been suggested such evidence has an important significance, but before going into this phase of the subject again let us examine a little more in detail the structure of these band fading records.

The steps in any one picture of Fig. 17 are, as we have said, snap-shots of the wave amplitude for successively different radio frequencies taken about a quarter of a second apart. The fact that the fifteen snap-shots used to build up a single picture are not taken simultaneously causes a skewing of the outlines when movement of the depressions as shown in Fig. 17 occurs. If, for example, we

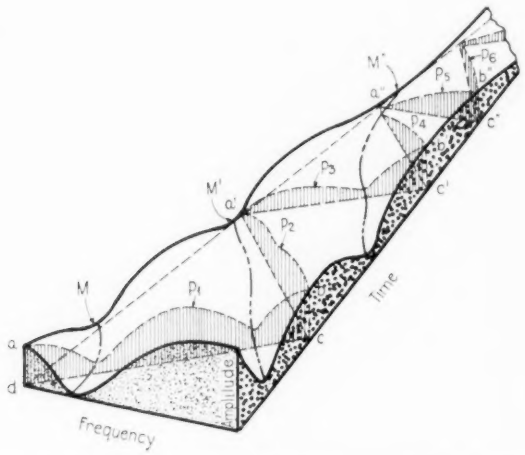


Fig. 18—Three dimensional diagram, showing the method of interpreting band fading records

were to take fifteen separate and successive snap-shots of a mountain through fifteen long vertical slits side by side it would be possible to combine the narrow sections so as to form a true picture of the peak. Now, if by some prodigious act of nature the mountain were shifted suddenly to one side and back again during the time we were taking the fifteen successive snap-shots through the vertical slits, the combination of them would form a profile quite different from that obtained when it was stationary. Or if it were simply moved steadily across the field of vision during the time the snap-shots were being taken one slope would be made to appear precipitous while the other would be leveled to a gentle grade in the finally built-up picture.

The character of this skewing, then, and its magnitude depend upon the rate at which the object being photographed in vertical sections moves, and the direction of the movement.

In Fig. 18 is shown an imaginary night-time band fading record in the "assembled" form. Since such a record contains frequency as a third dimension, in addition to amplitude and time as shown in the ordinary fading record, our simple fading curve has assumed the broader aspect of a surface, the selective fading making more or less parallel "valleys" running across it. The step-frequency system of recording the points amounts to photographing sections of this solid. The important point to be kept in mind is that these sections are *not perpendicular to the time axis.* If they were, the skewing previously described would not be present. By setting these sections up in their true relation to the time axis, however, and filling in to produce a continuous surface such as is shown in Fig. 18 the result is correctly represented. In order to make a detailed and accurate study of the band fading records, therefore, it is desirable to construct from the oscillograph sections the complete surface by the method suggested.

In Fig. 18 the trace of minima crossing the band is shown by M, M' and M''. Picture sections obtained as our recording apparatus literally moves back and forth across this frequency band are shown as $(a\text{–}b\text{–}c\text{–}d)$, $(b\text{–}c\text{–}a')$, $(a'\text{–}b'\text{–}c')$, etc. It will be evident that the section P_1, for example, will, in case a minimum is crossing rapidly, appear entirely unrelated to section P_2. When the minima run nearly parallel to the time axis (slow changes in transmission conditions) the successive pictures P_1, P_2, P_3, etc., will reveal their relation by direct comparison.

Actually to obtain frequency-amplitude sections perpendicular to the time axis in Fig. 18 would require the simultaneous transmission and reception of a large number of frequencies spaced at short intervals along the frequency axis. A more practical thought is to speed up the process and though this seems very simple at first consideration, it will be shown later to involve a particular kind of distortion which cannot be separated out as easily as the skewing encountered by the more deliberated method.

Now that we are familiar with the data, Fig. 19 showing, partially superimposed in vertical strips, the outlines of successive built-up pictures of the frequency traverse will be of greater significance. During the steady periods there appears within the 2,280-cycle band covered by these data approximately one complete cycle of selective fading. The lack of flatness in the audio-frequency-transmission characteristic of terminal apparatus has caused the suppression

of amplitudes toward the right side of these sections. Keeping in mind also the skewing inherent to this system of presentation during transient periods, we are able to trace the movement of minima, as illustrated previously in Fig. 17 which was taken from a different

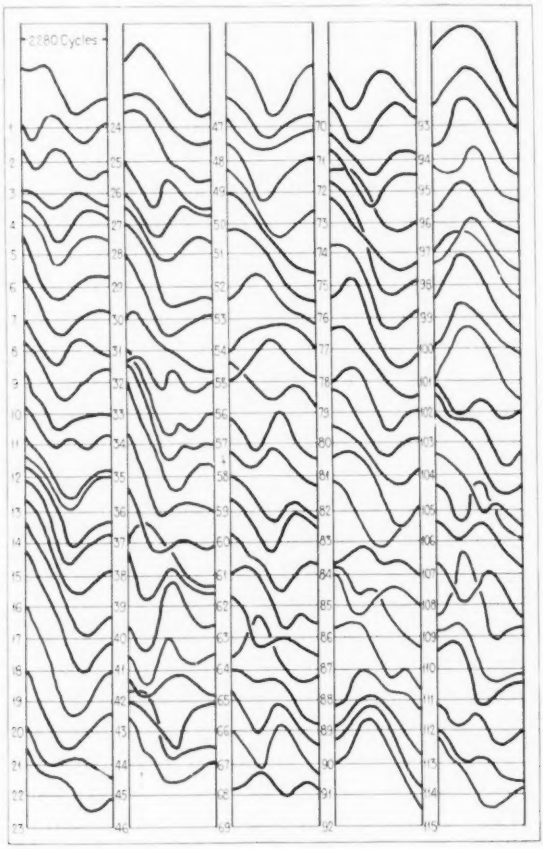


Fig. 19—Plotted curves, condensing a long band fading record so as to bring out the effect of selective fading

record. The relative position of these minima gives us an interesting insight into the nature of the night-time transmission path.

From records covering frequency ranges up to 4,500 cycles in width the positions of major minima along the frequency axis have

been plotted against time as in Fig. 20. The widths of the frequency bands covered in this case are indicated. This picture is essentially a bird's-eye view of band fading records such as are illustrated in idealized form by Fig. 18, the amplitude axis being perpendicular to the page. It reveals the presence of minima spaced at more or less definite frequency intervals, and suggests the presence of other

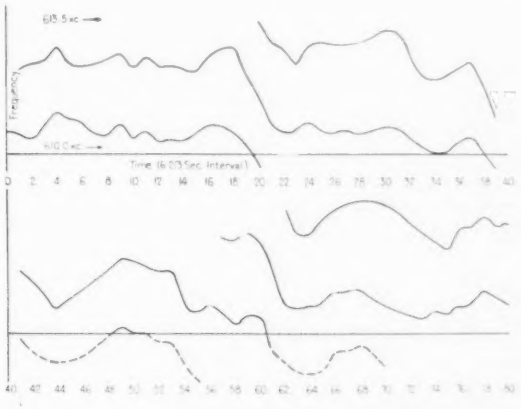


Fig. 20—Plotted curves which condense a long band fading record so as to bring out the frequency spacing interval of the selective fading

depressions in regular spacing beyond the scope of our pictures, for when one minimum slides out of sight another appears to take its place from the opposite side of the band. The minima traces shown in broken line were outside the record but were located by extrapolating the sections.

Other depressions of small amplitude appear to be superimposed upon the major changes but the present data appear inadequate to give reliable information concerning them. These minor depressions seem most evident during periods of rapid change.

The presence of these major minima in regular array bears a marked similarity to the familiar wave interference case in light and fits in very nicely with the theory detailed in previous paragraphs. Assume for a moment the simple case of two transmission paths producing such an effect and account for the difference in their lengths by presuming that one path follows more or less closely along the surface of the earth while the other seeks higher altitudes and in some fashion gets back to earth at the receiving station.

The mean frequency difference or spacing interval between successive minima for the records given in Fig. 20 is approximately 2,200 cycles. Therefore, the mean wave length difference in length of path from equation (5) is 277 wave lengths, or 136.5 kilometers.

It is evident that the errant waves following the second path must have been led a devious route. While this is about all the information which can be deduced directly from these data it is interesting to speculate further with the information along the lines of some of the theories which have been proposed to account for such wave deflections. For instance there is the Heaviside layer theory in which there is supposed to be a more or less well defined reflecting layer in the upper atmosphere. For this we would visualize our high altitude waves as proceeding in a straight line up to the layer, being reflected, and striking back to earth at the receiving station.

Since the distance from transmitter to receiver was 110 kilometers the length of the secondary path was 110+136.5 or 246.5 kilometers. By triangulation the height of the assumed reflecting layer may be determined as very nearly 110 kilometers or equal to the distance from transmitter to receiver, and the angle of incidence is 26.5 degrees.

As yet no positive information has been acquired concerning the variation of difference in length of two major night-time transmission paths with direct distance from the transmitter. If the path difference is due to reflection from an overhead layer, the expected relation by triangulation becomes quite simple.

$$\Delta d = \sqrt{\frac{y^2}{4} + h^2} - y.$$

When Δd is the difference in length of path, y is the direct distance and h is the vertical height of the layer.

An investigation of this relation would probably do much to prove or disprove the reflection theory.

At this point it is well to recall the results of earlier tests in which it was observed that single frequency waves separated by 1,000 cycles faded in approximately an inverse relation also indicating a spacing interval of about 2,000 cycles. The agreement of these earlier records is particularly noteworthy since about three weeks elapsed before the more detailed band fading records were made.

Fig. 20 shows a time variation in the frequency position of the minima which is explained as due to a variation in the difference of path length. If we indulge in further speculation along the line of layer phenomena we conclude that the reflecting layer is rising and falling. It is improbable that the whole layer would rise and fall

together so we conclude that undulations occur along its surface. These undulations in themselves would cause the length of path of the wave reflected toward the receiver to undergo a continual change. They would also introduce minor reflections from surfaces more distant than that responsible for the major effect which may be responsible for the more rapid, low amplitude fading which is usually superimposed upon the slow changes. Obviously, the character of the fading would in the event that it is caused by undulations along the reflecting layer, be determined by the amplitude and direction of movement over the surface.

If, on the other hand, we examine the possibilities of theories such as those proposed by Nichols and Schelleng, Larmor and others in which the action of free electrons in the atmosphere is invoked we might visualize the waves on the second path as following a curved trajectory. Or we might have the two sets of waves start off together, become split by double refraction and eventually come together again. Perhaps their planes of polarization will have been rotated. In fact it is possible to build up what appears, we must confess, to be a highly imaginary explanation in which the wave interference is accounted for not on the basis of any great difference in path length but by the assumption that the amount of rotation is such a function of frequency that a change of about 2,000 cycles adds or subtracts a complete rotation, and the further assumption that one set of waves has had its plane of polarization rotated through several more complete rotations than has the other. The synthetic possibilities are almost endless and we must wait upon further data more varied in character before the facts can be established. In the present investigation we have not attempted to determine the mechanism of the transmission medium except insofar as it could be inferred from the results of our tests which were aimed at finding out just how radio signals look after they have been subjected to a trip through this mechanism.

Returning to the solid band fading record illustrated in Fig. 18, let us form some conception of the appearance of this figure were it extended toward the much higher and lower frequencies using as a basis of this conception the supposition that the existing record is systematically distorted by wave interference. For a given rate of change in the physical difference in length of path, such as would be encountered in the simple reflection case, the rate of movement of the minima across the band fading pictures would vary directly with the frequency. Therefore, we can extend the narrow section shown in Fig. 18 to form a wide band fading record such as is shown in

Fig. 21, wherein we are looking down upon the distorted surface, the minima being traced by the light lines. Toward the short wave end of the band it is evident that a fading record for a single frequency represented, for example, by a section parallel to the time axis and perpendicular to the page, $a-a'$, would show rapid fading, while a similar record at the long wave end of the range as $b-b'$ would give slow amplitude changes. Such sections representing theoretical

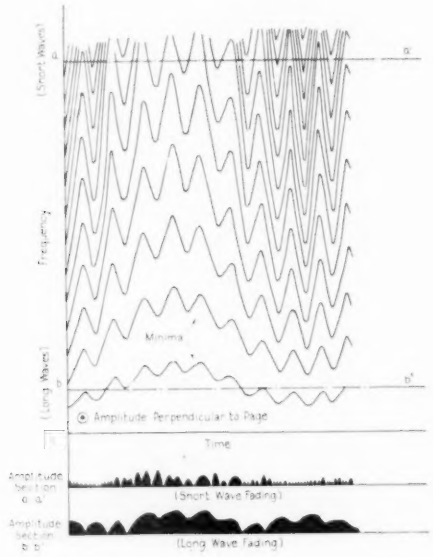


Fig. 21—Theoretical diagram obtained by extrapolating band fading records to show how the rapidity of fading might be expected to change with the wave-length

single frequency fading records are shown at the bottom of Fig. 21. The relative fading rates for long and short wave lengths as indicated by these idealized characteristics, are in accord with general experience

In describing the stepped-frequency method of obtaining band fading records allusion was made to distortion which might result from speeding up the process. Suppose that we were to use a very small rotating condenser in parallel with the main condenser of the transmitter oscillator for changing the frequency, and that this condenser were capable of changing the frequency sinusoidally about a mean value. Then we could represent the variation in frequency with time as is shown by the curve C_1 in (a) of Fig. 22. Now if the energy

transfer from transmitter to receiver takes place over two paths of different lengths one wave will constantly lag behind the other.

This lag may be measured as a time interval. In Fig. 23 are shown two waves, (a) and (b) of constant amplitude but with frequency

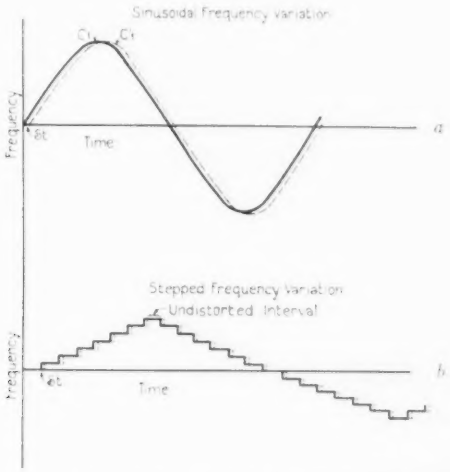


Fig. 22—Curves showing the relative effect of transmission time lag in sinusoidal and step-by-step methods of frequency variations

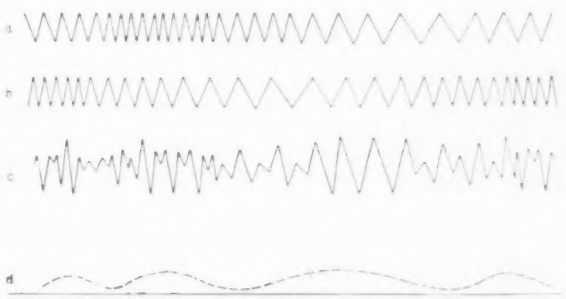


Fig. 23—Diagram showing the effect of frequency modulation

modulation. The wave (b) representing the indirect wave, it will be noticed, lags behind the direct wave represented by (a). The amount of this lag is determined by the difference in length of path and the transmission velocity. If we were to receive only one wave, as we should in the daytime, for example, we would find it to be a

constant amplitude field (providing the high-frequency characteristic of the receiver is flat over the range of frequency variation). But when two or more distinct paths exist, the combination at the receiver becomes complex. This is evident in curve (c) shown in Fig. 23 which is a direct summation of (a) and (b), and in (d) which is the envelope of (c). The amplitude is subjected to variations which did not exist at all in the original wave.

We might set up an *equivalent* effect right at the receiver by constructing two small local oscillators having the same characteristics as the transmitter oscillator. The two small rotating condensers would be driven by the same motor but the rotor of one would be shifted backward in phase relation to the other so as to simulate the case of transmission lag over the longer path. The relative frequency characteristics of the two may then be represented by curves C_1 and C_1' in (a) of Fig. 22.

The frequency of the signals arriving over devious paths at the receiver may be put in the form of an equation as,

$$F_1 = F_o + f \sin [r\ (t - d_1/V)], \tag{6a}$$

$$F_2 = F_o + f \sin [r\ (t - d_2/V)], \tag{6b}$$

wherein,

F_o = the mean frequency

f = one-half the total variation

$r = 2\pi$ times the frequency of rotation of the condenser

d = length of path

V = velocity of waves.

For a difference in length of path equal to 300 wave lengths at a frequency of 600,000 cycles per second, for example, the time lag of one wave behind the other will be equal to 300/600,000 second or 1/2000 second. The lag of one of the condensers behind the other in the "equivalent" case described above would be then for 30 cycles per second rotation of the condensers, 30/2000 times 360 degrees or 5.4 degrees. The lag of 5.4 degrees represents the lag of the condenser rotor so that the frequency lag will depend entirely upon the rate of change of frequency by the rotating condensers at any given instant.

Now to determine the resultant wave at the receiver we must know both amplitude and relative phase of the components arriving over the different paths. The amplitude will be constant, and we shall assume known, although it may actually follow slow changes with

attenuation or variations in length of path. The relative phase must be determined from equations (6a) and (6b). Knowing the frequency variation with time we may by integrating the following equation determine the phase relation at any time (t).

$$\Theta_1 = \int_0^t 2\pi F_1 dt, \tag{7}$$

$$\Theta_2 = \int_0^t 2\pi F_2 dt. \tag{8}$$

Substituting the general relation for F_1 and F_2 from equations (6a) and (6b) we have,

$$\Theta_1 = \int_0^t F_o + f \sin r\,(t - d/V), \tag{9}$$

$$\Theta_2 = \int_0^t F_o + f \sin r\,(t - d'/V). \tag{10}$$

Evidently the relative phase ($\Delta\Theta$) will be the difference between these two giving,

$$\Delta\Theta = \Theta_1 - \Theta_2 = 2\pi \int_0^t F_o\, dt + 2\pi \int_0^t f \sin r\,(t - d/V)\, dt \tag{11}$$

$$- 2\pi \int_0^t F_o\, dt - 2\pi \int_0^t f \sin r\,(t - d'/V) \tag{12}$$

which integrated reduces to the form,

$$\Delta\Theta = \frac{2\pi f}{p}\,(\cos r\,t - 1)\,(\cos r\,d'/V - \cos r\,d/V + \sin rt\,(\sin rd'/V - \sin rd/V). \tag{13}$$

The equation is not in itself very illuminating, but what it tells us generally is that if we represent two frequency modulated waves travelling over paths of different lengths to a distant receiver by rotating vectors, these vectors are constantly shifting their relative position. The magnitude of the shift at any instant is given by the varying angle $\Delta\Theta$. Due to a change in the angle included by the two vectors their resultant will undergo an amplitude change, the seriousness of which we will consider later.

Thus far in the discussion of frequency modulation by means of a rotating condenser we have assumed sinusoidal changes in frequency. The ordinary condenser departs considerably from such a performance. By considering the application of the integral equation for $\Delta\Theta$ to such a case it will be recognized that the relative space posi-

tions of the vectors representing the direct and indirect waves will be subjected to changes at every point where the slope of the frequency-time curve departs from a simple sine relation. The degree of distortion due to the presence of such irregularities may be considerable.

In Fig. 24 are shown some samples of "wobbled" carrier frequency records obtained at Stamford, Connecticut. For these records the

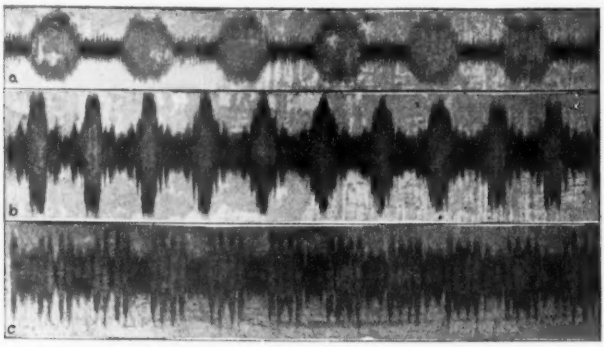


Fig. 24—Sample fast records showing distortion produced by intentional frequency modulation. *a* day record, *b* and *c* night records

carrier was wobbled at the rate of about 10 cycles per second. There is some uncertainty as to the range of frequency variation for these records although it was probably in the order of a few thousand cycles. By means of a constant frequency local oscillator the radio-frequency wave was stepped down in frequency to audio values which could be amplified and recorded.

The record (a) of Fig. 24 represents stable day-time reception. The record shows amplitude modulation due to the receiver characteristic alone. If the receiver were, as is desirable, capable of amplifying all the frequencies present in the received wave in the same ratio this record would be of constant width. In the subsequent examination of night records we must keep in mind the fact that the terminal apparatus is responsible for a certain part of the amplitude modulation. Its influence is readily recognizable.

The night-time records shown in (b) and (c) reveal a distinct distortion of the envelope aside from that present in the daytime record. Peaks appear and disappear within time intervals sometimes as short as a fraction of a second.

The record in Fig. 25 represents a slow picture of the changes shown in (b) and (c) of Fig. 24. If these wobbled frequency waves are studied carefully it will be noted that where a single peak stands at one moment there gradually comes in view another as if it were sliding from behind the first. The cycle length being about 1/10-second we may get some idea from this series of the rate at which the changes

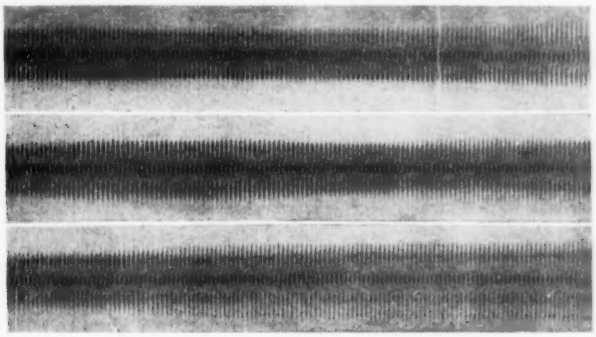


Fig. 25—Sample slow record showing distortion produced by intentional frequency modulation. Night record

take place. The presence of so many peaks in these records is attributed in part to the fact that the rotating condenser used gave a frequency change which was far from a simple sinusoidal relation.

Let us now return to the stepped-frequency method of obtaining the band fading pictures and ascertain why it has certain advantages. In (b) of Fig. 22 is shown the "equivalent" characteristic for the stepped condenser. During 1/2000 of a second (for the conditions so far assumed) in each step distortion may occur due to transient conditions, but during the remainder of the quarter second assigned to each step (for the records so far taken) a steady state is reached. Thus, theoretically, distortion occurs only during about 1/500 of the step interval. In (b) of Fig. 22 the lag is greatly exaggerated for purposes of illustration. This means simply that we have maintained constant frequency for a sufficient length of time to establish, before taking our picture, a fixed interference condition over the region including transmitter and receiver at least.

Daytime Field Strength Distribution

Thus far we have been dealing with the unstable phenomena of night-time transmission. Our interest has been directed almost

entirely toward variations with time. While the presence of wave interference has been detected, and the movement of this interference effect across the frequency band has been recorded, little effort has been made to form a picture of such interference in its space relation. A discussion of similar stable, daytime phenomena is therefore not out of place, and particularly so in view of an evident relation of the fickle nocturnal interference phenomena to the steady states which follow the appearance of daylight.

In a previously published map of field strength distribution in New York City,* it was indicated that the congestion of high buildings

Fig. 26—Map showing location of radio obstruction on Manhattan Island as determined by the intersection of lines between various transmitting points and their corresponding shadows

just below Central Park cast a heavy shadow. More recently it has been determined from observations on a portable transmitter, set up at various points, that this building center is a consistent performer. The position of this obstruction is determined in Fig. 26 wherein only partial contours from maps for the indicated sites are given to prevent confusion. The intersection of these lines from transmitter to shadow, falls at approximately 38th Street in the vicinity of Sixth Avenue.

The dissipation of wave energy at such a point is probably the composite effect of many adjacent structures. Fig. 27 gives an elementary idea of how this can occur. The structures filling in

* See footnote 1.

each block are, of course, very well connected electrically by means of pipes, cables, etc., with those of adjacent blocks. Between each oscillating circuit (which is pictured as consisting of two buildings with earth connections) there exists a coupling which binds the whole system together more or less flexibly. Thus the obstacle offered by

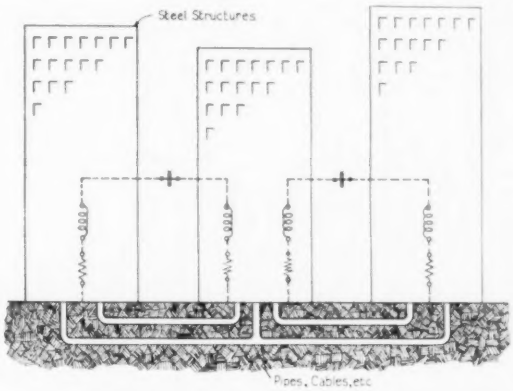


Fig. 27—Idealized picture of equivalent electrical circuit characteristics of high buildings

a group of buildings might be of a selective nature, and evidently its frequency characteristic may vary with direction.

Such an aggregate would, in addition to absorbing wave energy, produce a change in velocity or a refraction of the wave front. Some indication of such an effect will be discussed later. Before leaving the subject of shadows, however, let us get a physical picture of their significance.

From the transmitter a wave front expanding outward and upward encounters an obstruction which we shall assume is near the earth plane. The net result of this encounter is a weakening of the wave over an area near this plane, and probably a distortion of the energy-bearing fields. We might then imagine this shadow to be a tunnel-like region extending along the earth beyond the obstruction, and as having definite vertical as well as horizontal limits.

The aerial photograph of Manhattan and adjacent territory, shown on Fig. 28, will give a fairly clear idea of the conditions close to the transmitter. The major obstruction, the location of which has been previously described, is shown in its relation to the line of transmission toward the Riverhead and Stamford testing stations.

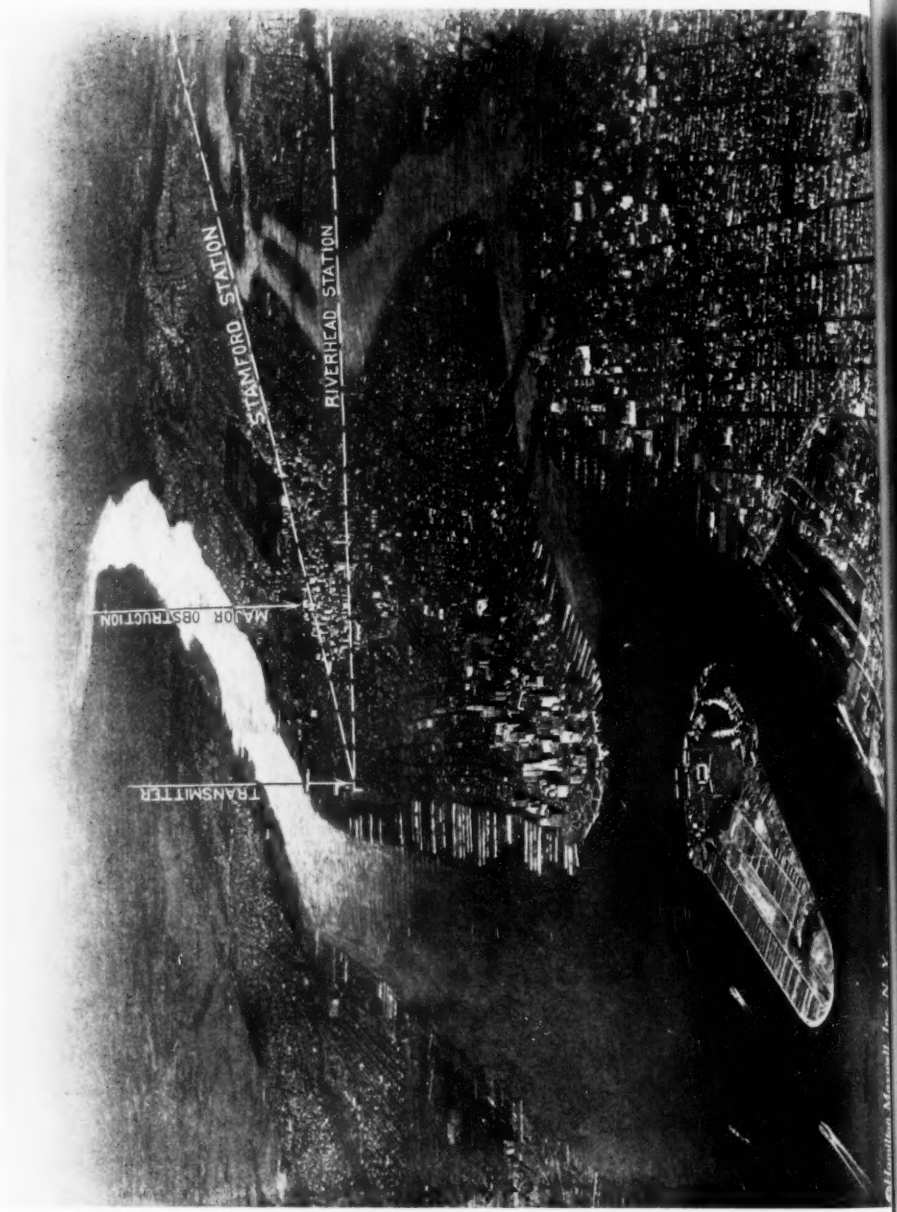


Fig. 28—Aerial photograph of Manhattan Island showing locations of transmitting station and obstructing high building area

Such barriers to wave travel, situated within a short distance from the source, seem, as we might expect, to have a more extensive and serious influence upon effective broadcast distribution than similar obstructions at greater distances.

It will be noticed that the obstruction falls very nearly upon the direct line from the transmitter to the Stamford testing station. This will also be evident later after an understanding of Fig. 29.

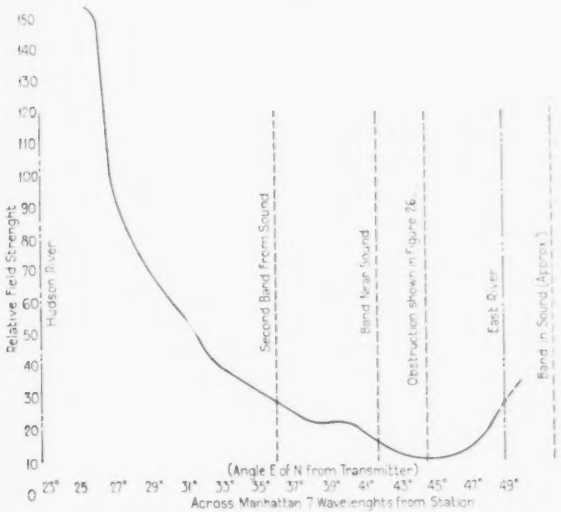


Fig. 29—Cross section of radio shadow caused by high building area

wherein the position of the "Band Near Sound" represents also the bearing of the Stamford station. The Riverhead station is not directly in line with the major obstruction.

In certain sectors of the field strength contour map for station 2XB there appears to exist a kind of wavy displacement of the contour lines forming a partial pattern of peaks and depressions side by side. In general, this pattern must be differentiated from an ordinary shadow area. A remarkable example of this sort of field distribution is shown in Fig. 1 which is one section of a field strength survey made for station 2XB. These contours are based entirely upon daytime measurements, and represent a condition which is stable throughout the daylight period. Considerable difference in signal level is apparent within short distances across the direction of wave propagation. Two pronounced low signal channels extend ap-

proximately north-east across this region. These shift with change in frequency of the transmitted wave. Fig. 30 illustrates the space relations for such a movement. The full line curve shows a partial cross section of the contour map of Fig. 1 taken along a line approximately perpendicular to the direction of transmission 110 wave lengths from the transmitter. This represents relative field strength values for

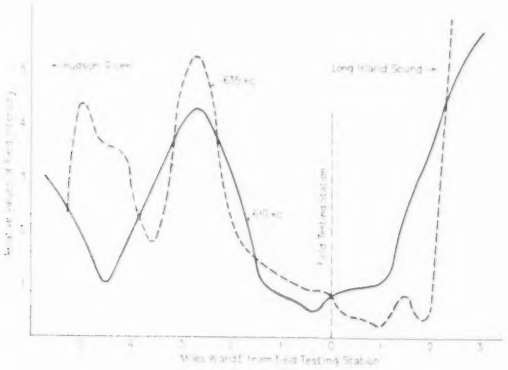


Fig. 30—Cross-section of wave interference pattern showing change with frequency

610.0-kilocycle radiation. When the frequency is raised to 635.0 kilocycles, there occurs a movement of the peaks and depressions as is shown by the broken line of Fig. 30. Apparently the increased frequency causes these channels to be crowded together.

If we take sections of the field strength contour pattern in Fig. 1 and examine carefully the relative amplitude of peaks and depressions represented by these wavy lines we shall find that the ratio of field strength of the peaks to that in the depressions increases with distance from the transmitter. That is, the channels become more sharply defined as we move away from the transmitter. This ratio is shown approximately by the curves of Fig. 31. If these peaks or depressions were simple shadows they would maintain their relative values at a distance from the source or even tend to "heal" causing the ratio to fall rather than rise as is actually the case.

Within 14.4 wave-lengths (7.1 km.) of the transmitter the pattern, so apparent beyond 30 wave-lengths, merges into one deep shadow a cross-section of which is shown in Fig. 29. The abscissa of this curve is in degrees measured from the transmitter so that the center of the two most distinct low field strength channels extending north-

east may be inserted with their true radial relation. The two most evident in Fig. 1 are shown to be west of the line extending from transmitter through the center of the obstruction located in Fig. 26. The presence of Long Island Sound east of the geometrical center of the shadow has made an extensive survey of this section imprac-

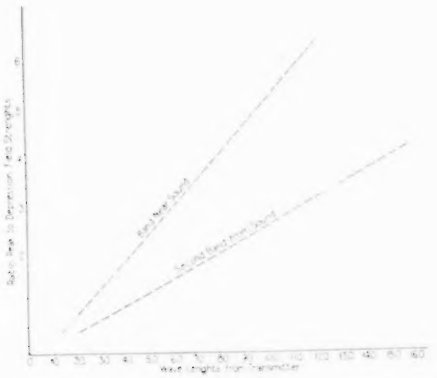


Fig. 31—Plot showing intensity of definition of wave interference pattern

tical. However, a single section taken across the Sound at about 90 wave-lengths from the station shows quite unquestionably the presence of a low channel about as indicated to the right of the obstruction designated in Fig. 29.

We have, therefore, a deep shadow with a more or less orderly array of maxima and minima within its limits. These maxima and minima grow more distinct at a distance from the transmitter, contrary to what we might expect for ordinary shadows. Furthermore, we find that they move as the frequency is changed. These facts lead to the belief that the phenomena in question are due to wave interference such as has already been described in connection with night-time fading, but characterized by very much smaller path differences. This daytime interference condition is fixed while we have seen that the nocturnal patterns appear to wander continually. To explain this more in detail let us return to the shadow and consider the phenomena which might accompany it in a little more detail.

The study of light has made available much information concerning the subject of wave interference. It is known, for instance, that the edges of shadows are not sharply discontinuous changes from light to darkness, but that a series of dark and light bands, called

diffraction fringes, are interposed between the full light and full dark areas. In our radio case the distance from the source to the obstruction and the dimensions of the obstruction are both very much smaller, in comparison with the wave length of the radiation, than for any ordinary case in light, but apparently the phenomenon is of the same general nature. By applying the ingenious principle of secondary sources used by Huyghens we might theoretically determine the distribution of the field beyond an obstruction placed in the path of the advancing radio waves. The basis of this principle is

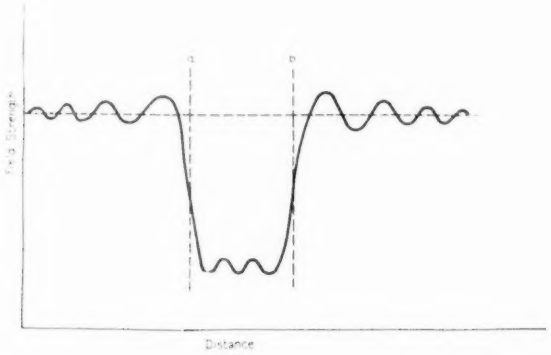


Fig. 32—Theoretical cross-section of radio shadow and associated wave interference pattern

the assumption that each elementary part of the advancing wave may be considered as a tiny transmitter. The effect at any point behind an obstruction, therefore, becomes the resultant effect, considering phase as well as amplitude, of the waves from all these miniature sources.

In Fig. 32 the region between vertical lines (a) and (b) represents the geometrical limits of the cross-section of a well defined shadow taken some distance behind the obstruction. An analysis of the resultant field using Huyghens' construction would show variations in intensity somewhat as represented by the full line. In other words the shadow will not be distinct but will have alternate maxima and minima within its geometrical limits and similar variations beyond the edges.

It is very likely, of course, that even in case the foregoing speculative analysis of the contour pattern extending north-east of 2XB is fundamentally correct, a great many other influences than that

of obstruction enter into the final field distribution. Relative attenuation of water and land appear to influence the distribution considerably though not as definitely as do steel structures close to the transmitter. Distinct minima appear both on the Hudson and on the Sound along radial lines extending from the transmitter.

Probably refraction of the wave front in passing across shore lines also enters into the shaping of this pattern.

Perhaps as good an elementary picture as any of the phenomena causing these patterns is that of a "dent" produced in the wave front by an encounter with a portion of New York City's impressive skyline. Since radio waves travel in a direction perpendicular to the plane containing the electric and magnetic fields, opposite sides of this "dent" would cross over one another with the result that an interference pattern would appear beyond the obstruction. An analogous situation exists when a water ripple passes a cluster of marsh grass which, damping its motion and retarding its progress causes part of the advancing front to converge and cross beyond the obstruction.

There is evidently a relation between day patterns such as have been discussed and night-time conditions. Just what this relation is offers some further opportunity for conjecture. In the first place quality distortion in transmission at night was, as previously explained, observed over parts of the region covered by the pattern shown in Fig. 1. The worst distortion seemed to be somewhat associated with the low field strength regions in this daylight survey. The distortion seemed also to be worse along the low channel extending in the direction of New Canaan, Conn., and beyond the 100-wave-length circle. It was particularly bad at a distance of some 140 wave-lengths from the station along this low channel where the field strength became so low in the daytime as to be unmeasurable with the set employed for the work. Accompanying the poor quality were fading and marked directional shifts.

Quality distortion though not so consistently severe at the Riverhead station as in the vicinity of Stamford was at times easily detectable by audible tests. Due to rapid attenuation of the radio waves traveling from the site of 2XB across Manhattan and the length of Long Island the field strength around Riverhead is generally low with higher levels north and south on the open waters of the Sound and Ocean respectively. Night-time fading at this point was representative of the variety which is usually found at distances of approximately one hundred miles from a broadcast transmitter.

The situation at Riverhead appears to be somewhat the same as

that which may exist over a large part of the broadcast area at a distance from the transmitter, while in the Westchester region we have an extreme and rather special circumstance. Field strength surveys have shown that there are indications of a daytime interference pattern over the Riverhead area but this pattern, such as it is, appears to be irregular and to lack the definition which makes the Westchester pattern so remarkable.

On the basis of the Westchester data alone we might build up a theory to the effect that night-time shifts of the stable daylight pattern were in some way responsible for quality distortion following the departure of daylight. Such a thought applied to the Riverhead case does not seem so reasonable since here the pattern is about one-quarter as distinct in terms of the ratio of maxima to minima values as the Westchester pattern. If, however, we presume that quality distortion may be expected in areas where daytime signals arrive *considerably attenuated* or so interfering as to simulate such an attenuated condition both situations are satisfied. After a consideration of the evidence at present available, such a conclusion seems attractive; that is, a daytime wave interference pattern alone is only an agency in night-time quality distortion in so far as its minima in combination with the general shadow effect are responsible for a low signal *directly* transmitted. Perhaps, in other words, the daytime field strength is a measure of *direct* night-time transmission, there existing in combination with this direct path at night a second, variable route of greater effective length. Probably close to the transmitter the "direct wave" is large compared to the "indirect" but shadows or interference may materially modify the ratio.

Night Distribution of Field Strength

By receiving simultaneously at several points the signal coming from a distant transmitter, it ought to be possible to detect the movement in space of these interference bands we have been discussing. The question immediately arises as to how far apart these distributed receivers can be placed without giving us an entirely discontinuous and misleading picture. For the first step toward recording space variations, in the vicinity of the Riverhead testing station, the receivers were spaced 1/16 wave length (30.5 meters), as illustrated in Fig. 33. It is necessary in making such determinations to transmit a single radio frequency, since we have already found that the interference bands for one component of a modulated wave are likely to be in a different position than those for another.

In order to receive and record the radio frequency wave it is, as has already been shown, convenient to use a local oscillator to beat it down to audible values. Since several oscillators for the separate sets are likely to produce mutual interference a common one was

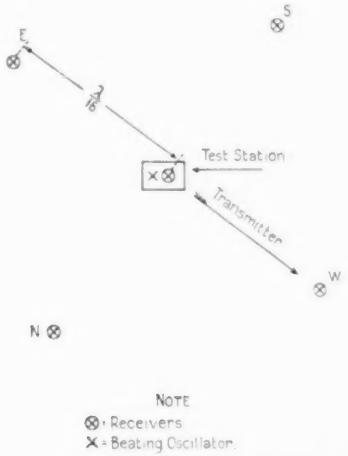


Fig. 33—Diagram showing space relation of receiving sets for special test

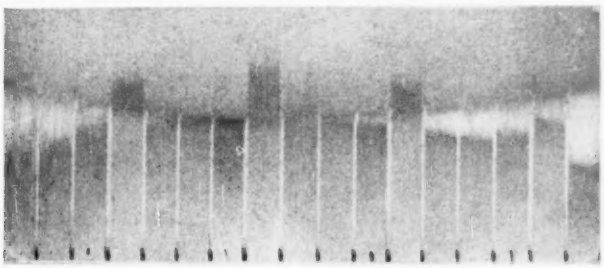


Fig. 34—Sample single-frequency fading record from spaced receiving sets

employed. This beating oscillator was situated at the testing station and the receiving antenna at this point was used as a radiator. In order to prevent overloading, the local receiver, the coupling to the receiver input coil was balanced to give a minimum of the local signal.

Fig. 34 is a sample of the record obtained. The continuous shadow band at the top represents the local receiver output. One oscillator

element was used for the other four receivers, their signals being recorded successively by a commutating device. Incidentally the interaction between these receivers was checked by observing the output of any one, while changes were made in the tuning of the others. The antenna was, however, so nearly aperiodic that no recognizable distortion or reradiation phenomena could be detected.

Fig. 35 illustrates compactly variations recorded by the oscillograph records (of which Fig. 34 is a sample), for a representative period of about five minutes. Even within the dimensions of 1/16

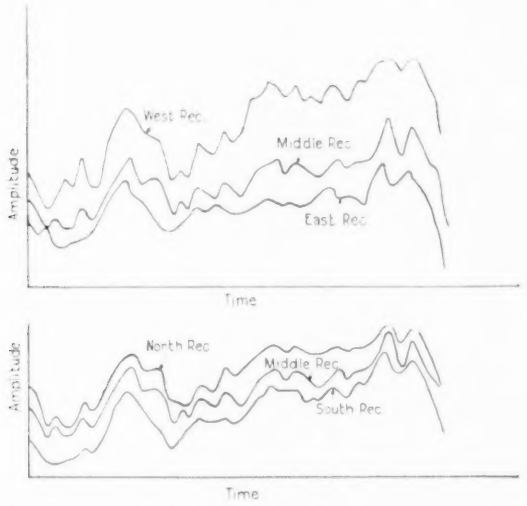


Fig. 35—Curves showing single-frequency fading on spaced receivers, condensed from long record

wave length there appears to exist transient field strength gradients in the direction of transmission. This is shown by a change in relative values, in the upper set of curves which represents field strength at points 1/16-wave length apart in the direction of transmission. The deviation is particularly noticeable in the relation between values for the local receiver and the "West receiver" which is in the direction of the transmitting station.

The lower set of curves, representing similar values across the line of transmission are much more nearly parallel. From the data so far obtained for the Riverhead testing site, it seems that transient nighttime field strength gradients are more generally evident in the direc-

tion of transmission than perpendicular to this direction. Upon these limited data one might be tempted to predict the presence of interference bands across the line of transmission.

The above discussion concerning space relation of field strengths has been included merely by way of contributing an additional bit of evidence to the theory that the erratic type of fading ordinarily

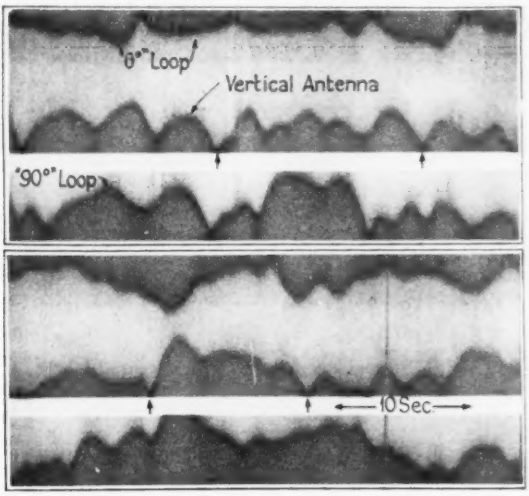


Fig. 36—Single-frequency fading record from vertical antenna and two-loop antenna crossed at right angles

experienced at night time is due to wave interference. The picture is very small in terms of wave lengths but considering its content, its very limits seem to imply wave interference rather than attenuation alone.

In connection with the wave interference theory thus far suggested as responsible for a major part of fading Fig. 36 is introduced as added evidence. The middle record of this group represents amplitude changes in the night-time reception of a carrier wave upon a vertical antenna. The upper and lower records represent the same for two loops turned at right angles to one another in the horizontal plane. By daytime tests the interaction of this combination was found to be negligible. Night-time fading recorded simultaneously for these three separate receivers occupying as nearly the same point in space as was possible, show that a high amplitude

signal may be coming in on both loops while the vertical antenna pick-up approaches zero. Several points of this kind are marked by arrows below the middle trace in Fig. 36.

There are at least two simple possibilities which might account for these relations. In case the wave approaches the receiving point from directly overhead, the vertical antenna would receive a "zero" signal while the loops would pick up an amount depending upon the state of polarization. If this be true, the records indicate a very rapid shift from the vertical direction of reception since the antenna minima are short lived most of them lasting at best a small fraction of a second.

On the basis of wave interference it is apparent that two waves approaching the receiving point in a 90-degree space phase relation and 180 degrees out-of-time phase could give a maximum signal on the two loops while that received on the vertical antenna was a minimum.

A compromise between these two viewpoints is probably a better guess than either one of them taken alone. That is, the existence of minima on the vertical antenna at the same moment that a strong signal is coming in on the loops is perhaps due to the interfering combination of waves having components in both the vertical and horizontal planes.

Quality Distortion

So far the data shown have been limited to the results of observations taken on special forms of transmission which are simplified for the purpose of clearly exposing the basic facts. We wish now to consider some of the more practical aspects of signal distortion. The first test which we made at our field test station was to record on slowly moving photographic paper tape and on the high speed film, the detected audio signal which resulted when the transmitter was modulated by a pure 264-cycle tone.

Fig. 37 is a sample of the general type of audio signal record obtained and Fig. 38 shows copies of the wave shape of the received signal, at particular times corresponding to the numbers of the oscillograms on the records in Fig. 37. The abrupt displacement of the timing trace indicates the point on the long record at which the snap-shot oscillogram was made. A peculiar characteristic of these records is the dark shadowy lines weaving back and forth through the band recording the complete signal. These dark lines correspond to the kinks in the wave shape shown in Fig. 38. As explained before, the darkening of the record is caused by the greater quantity of light

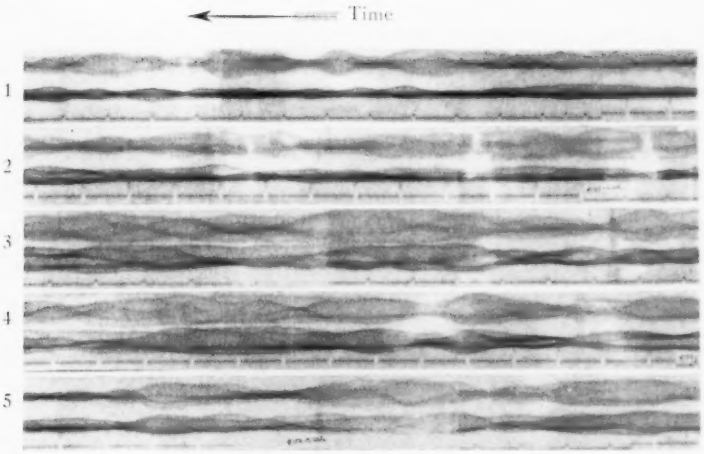


Fig. 37—Slow record of signal detected from tone modulated transmission showing the night-time distortion. Made at Stamford, Conn., May 15, 1924, 2:25 a.m. Upper trace signal from vertical antenna receiver and lower trace signal from loop antenna receiver, timing marks 2.6 seconds apart

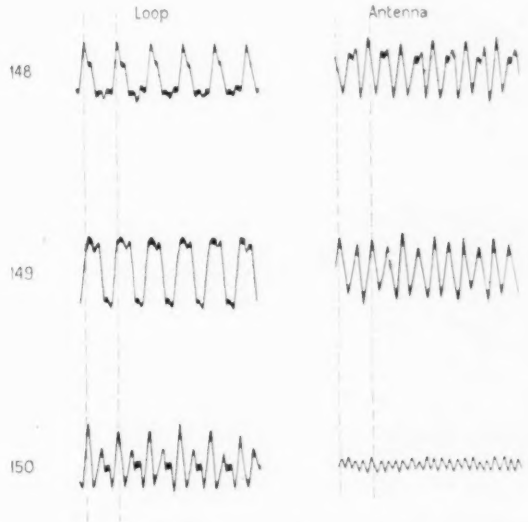


Fig. 38—Wave form of signals corresponding to numbered positions indicated on strips 2, 4, and 5, Fig. 37

affecting the record at these peak points. At the same time these observations were made, the wave shape of the signal rectified from the antenna current at the transmitter was recorded by an oscillograph. These oscillograms showed the signal to be free from distortion at the transmitter.

The weaving of these shadowy traces together with their width gives a record of the change in phase and amplitude of the irregularities in the wave shape of the signal. Although the wave shape of the signal is continually changing, it persists in substantially the same form for a great many cycles. Thus the record shows that, in the transmission of this simple tone modulated signal from the transmitting to the receiving antenna, it has been so modified that entirely new frequencies appear at the receiver. This receiver was shown by local tests to be free of any appreciable distortion within itself. While these new frequencies look like harmonics of the modulating tone in the snap-shot record it is obvious from the slow record that they are not true harmonics but that they differ from the harmonics by a very small amount and are incommensurable with the modulating tone since they undergo progressive but irregular phase changes with reference to it.

These records represent in a nutshell the signal distortion problem as it first presented itself to us. Our work then consisted in raveling out the complicated relations so that their nature could be ascertained and a theory of the causes established. In this paper, in the interest of clarity of presentation we have departed considerably from the actual order of the experimental work but at this point perhaps the actual order is best to follow for a moment.

With such a weird-looking distortion to analyze, and if possible eliminate, our first thought was as to whether the terminal apparatus might not involve unrecognized peculiarities which would be a contributing cause. Local tests and daytime tests of the receiving system absolved it from doubt and attention was focussed on the transmitting apparatus.

It was suspected that present day radio telephone transmitters leave something to be desired in regard to what we may call, for lack of a better term, their dynamic frequency stability. A very large percentage of the transmitters in use throughout the world today produce amplitude modulation of the carrier by the action of modulating tubes directly upon an oscillating tube circuit. It is to be expected that the cyclic changes in circuit conditions occurring at the modulating frequency will have some cyclic effect on the absolute frequency of the carrier and that this effect will be in the nature of a

wobbling or rapid shifting back and forth in frequency of the amplitude modulated carrier. In other words the carrier and side-bands, without change in their relative frequencies, would be subjected to "frequency modulation."

This sort of thing should be clearly distingushed from the slow wandering of frequency which, for instance, causes beat notes between carriers of different stations to drift gradually in pitch. What we have called "dynamic instability" is so rapid (being governed by the cyclic variations of the modulator) that it is difficult to observe by any aural method. Since the transmitter being used for our tests

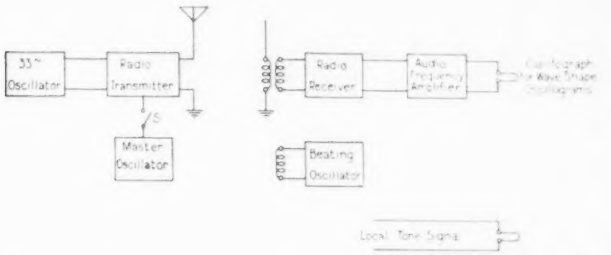


Fig. 39—Diagram of system used to measure frequency modulation

was a member of this almost universal class which employs modulating elements directly associated with the oscillator elements we determined to study this aspect of the transmission.

The following test was made to find out the extent of the frequency variation during the period of the modulating cycle. A schematic of the testing circuit arrangement is shown in Fig. 39. The plan was to modulate the carrier with 33 cycles, a tone so low in frequency that it would not be efficiently transmitted through the audio frequency amplifier connected to the output of the radio receiver. Then upon beating the received modulated carrier signal down to a frequency of about 1,000 cycles, an oscillogram of this signal would show a 1,000-cycle signal with a 33-cycle modulation in amplitude. Frequency modulation, if present, should then be easily discernible from the record. This experiment was made for daytime transmission and oscillograms (A) and (B) shown in Fig. 40 were obtained, one with the frequency of the beating oscillator greater than the carrier frequency, and the other with the beating oscillator frequency less than the carrier frequency. Both of these oscillograms show by the change in the frequency of the beat note signal

that frequency modulation occurs in the transmitter circuit. The frequency change is very apparent on the oscillograms when the lengths of one cycle at maximum and minimum amplitudes are compared. The reality of the effect is demonstrated in the two records, which by their difference show the reversal of the increased and decreased frequency points with reference to the modulation cycle when

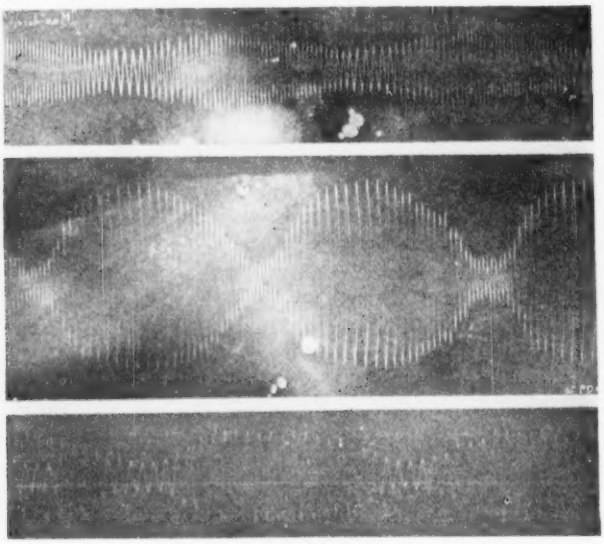


Fig. 4.0—Oscillograms showing frequency modulation accompanying amplitude modulation

the beating frequency is moved in frequency from one side of the carrier to the other.

The next step was to determine to what extent a stabilization of the carrier frequency to stop frequency modulation would affect the distortion of signals. True, master oscillator transmitters capable of giving the desired stability are not a new thing in the art. Several such transmitters were built by the Western Electric Company some years ago and used successfully in ship-to-shore radio telephone experiments[2] in which frequency stability was of considerable importance. To modify the ordinary broadcasting transmitter to in-

[2] See Fig. 1 and accompanying discussion in: Radio Extension of the Telephone System to Ships at Sea by H. W. Nichols and Lloyd Espenschied Proc. I. R. E., Vol. II No. 3.

clude this feature involves major mechanical changes and in order to provide a suitable arrangement for these tests the Bell Telephone Laboratories engineers merely added to the existing transmitter at station 2XB a temporary separate oscillator and high-frequency amplifier which could be connected to drive the oscillator tubes of the set as amplifiers. That this was free from frequency modulation is seen by comparing (C) of Fig. 40 with (A) and (B).

The transmission tests carried out with this arrangement yielded highly satisfactory results as indicated by a comparison of Fig. 41

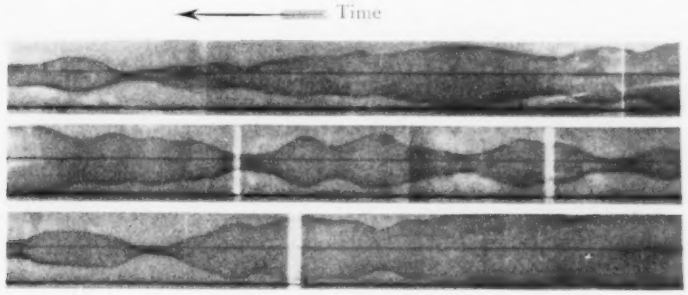


Fig. 41—Slow record of signal detected from tone modulated transmission with stabilized carrier showing reduction in distortion. Made at Stamford, Conn., Oct. 10, 1924, 3 a.m.

with Fig. 37. Fig. 41 like Fig. 37 is the detected result of a signal which started from the transmitter as a pure tone modulated signal, but it shows that much of the wave form distortion has disappeared, there remaining only a residuum which characteristically appears at the lower amplitudes of the signal. The probable cause of this residual effect will be discussed later. Tests of speech and music were concurrent with these findings. Using the normal transmitter, night-time transmission as received at the test stations was seriously distorted. When the stabilizing arrangement was employed this distortion was apparently eliminated except at the minima of fading.

Having arrived then at this practical result we wished to make further confirming tests, and tests to determine the whys and wherefores of the result. We have already detailed the more basic of these tests in previous sections of this paper and are now ready to consider the practical distortion records more carefully and to build up a theory to explain them.

The records shown in Fig. 42 are similar to the records in Fig. 37. They are shown here to illustrate the difference in the characteristics

of the wave form distortion variation that occurs from day to day. All these records were made at Stamford, Conn.

Strips 1 and 2—May 15, 1924—4:30 a.m.
Strips 3 and 4—Jan. 23, 1925—5:30 a.m.
Strips 5 and 6—Jan. 24, 1925—6:15 a.m.
Strips 7 and 8—Jan. 24, 1925—8:00 a.m.

There is a marked difference in the records obtained on January 23 and 24, which were made at the time an effort was being made to determine the effect of the solar eclipse on radio transmission. The peculiarly

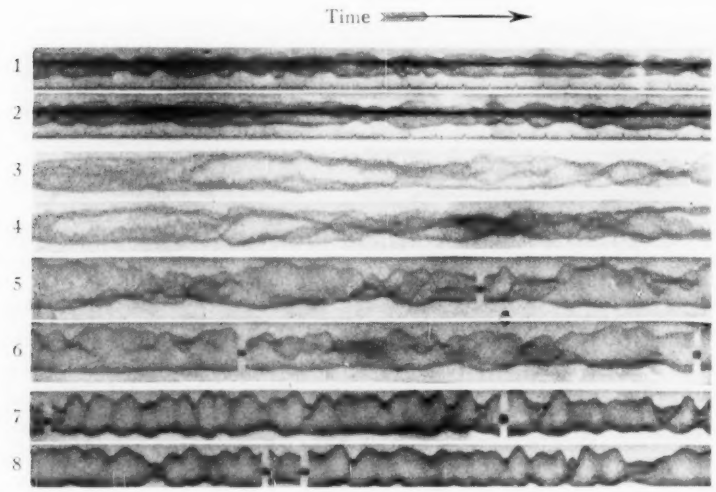


Fig. 42—Slow record of signal detected from tone modulated transmission taken on different days showing the changes in the character of the distortion

twisted appearance of the record obtained on January 24 is not very common in the records obtained. Most of the records have characteristics similar to those shown in Fig. 37. In the January 24 records there is a marked change in the characteristic configuration of the variation.

In order to obtain a record of the amount of wave form distortion resulting from frequency modulation present in the detected audio signal the circuit arrangement shown in Fig. 43 was used. This circuit was designed to analyze the wave form distortion when a 250-cycle

signal was used to modulate the carrier. Special precautions were taken to obtain a pure 250-cycle modulating tone. The wave shape of the signal detected from the carrier at the transmitter was frequently checked by observations with an oscillograph. The signals detected from the antenna current at the transmitter, both for the normal

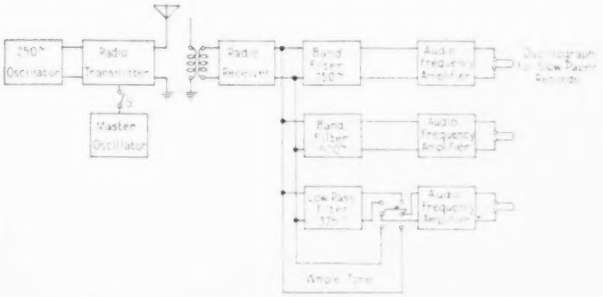


Fig. 43—Diagram of system used to obtain "Harmonic" analysis distortion records

transmitter with frequency modulation and for the stabilized carrier transmitter, were practically simple sine waves. The output circuit of the radio receiver was connected to a group of filters designed to transmit narrow bands of frequencies straddling the harmonics of 250-cycles.

While below we have referred to the frequencies passing these filters as "harmonics" it should be borne in mind that they are not necessarily *true* harmonics since they deviate very slightly from the true harmonic relation. The purpose of the test was to procure a record which would show at a glance the presence or absence of wave form distortion.

The input circuits of the filters were connected in parallel and the output circuits separately connected to the audio amplifiers arranged to operate the oscillograph elements. The input of one amplifier was arranged so that it could be switched either to the output of the filter passing 250-cycles or the output of the radio receiver. In this way a record could be obtained of either the whole tone from the receiver or only the 250-cycle component.

In Fig. 44, Strip 1 is a harmonic analysis record of the audio tone detected from the carrier and both side bands, transmitted with a stable carrier frequency. Strip 2 is a section of a record made a few minutes later when an unstabilized carrier was being used. On this record the lower trace is the 250-cycle component, the center trace the

500-cycle component, and the upper trace the 750-cycle component. The upper and lower traces have their zero lines at the edges of the strip. This record was made at Riverhead, L. I., April 30, 1925, at 3:33 a.m. Strip 2 is a section of a record made a few minutes later when an unstabilized carrier was being used.

The gain in the audio amplifiers connected to the outputs of the filters was adjusted to give nearly uniform transmission through the

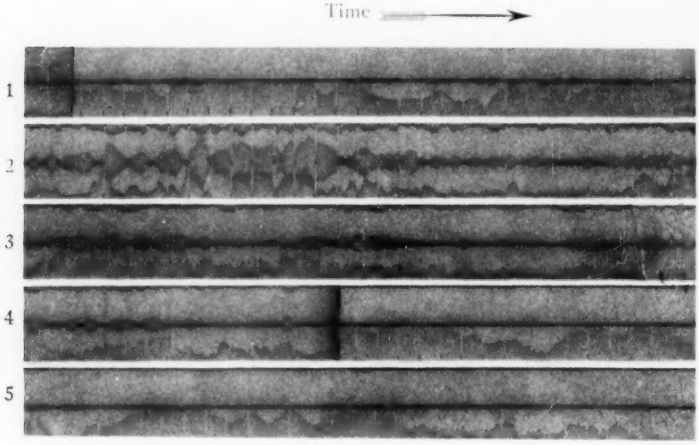


Fig. 44—Slow record made with system diagrammed in Fig. 43. Contrasting the distortion of detected tone transmitted by stabilized and unstabilized carrier frequency

receiving and recording apparatus for the frequencies recorded. Hence in these records the relative amplitudes of the fundamental and harmonics of the signal are directly comparable.

Strips 3, 4 and 5 in Fig. 44 are taken from a record made for the purpose of obtaining a comparison of the wave form distortion sustained by the detected audio signal transmitted by the normal transmitter with frequency modulation present and by a stable frequency transmitter. In each strip the lower trace is the whole tone from the output of the radio receiver, the middle trace the second harmonic (500 cycles) and the upper trace the third harmonic (750 cycles). Strip 3 and half of Strip 4 give the record obtained when the normal transmitter was used, and the remainder is the record obtained when the modified transmitter was used. There was a few minutes' difference in time between the ending of one transmitting condition to the beginning of the next during which the master oscillator control was switched

on at the transmitter. The receiving circuit was not changed during the making of this record, so that the results obtained from the two transmitters are directly comparable.

The record of the signal from the normal transmitter shows an abundance of second and third harmonics, at times equal in amplitude to that of the whole tone signal. The latter, of course, includes these harmonics. It will be noted also that dark line shadows run through the trace of the whole tone, indicating the presence of the wave form distortion. The signal from the stable frequency transmitter as shown by the record is practically free from wave form distortion. The trace of the whole tone is also free from any dark lines which would

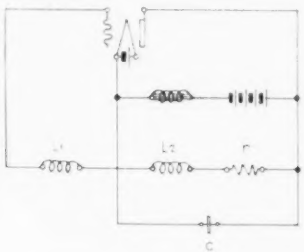


Fig. 45—Diagram of an oscillator circuit

indicate wave form distortion. This record is substantial evidence that a great deal of the wave form distortion may be eliminated when the carrier is stabilized. However, the selective fading still remains.

The selective fading we have already explained more or less satisfactorily and we find that it does not materially affect the wave form of audible frequencies transmitted by a modulated stabilized carrier unless its changes are more rapid than any we have recorded. The crippled state of originally perfect tone waves after they have been transmitted by an unstabilized carrier, we have just observed. Now let us consider the possible causes of this difference. The carrier stabilization referred to here, may we repeat, is not stabilization against slow variations in frequency from second to second or from hour to hour but rather against rapid variations within the cycle of the modulating frequency.

The reason for such changes over the modulating cycle is that the variation of the impedance of a vacuum tube across the oscillating circuit necessarily causes a variation in the nature period of the oscillation. As a simple case, the circuit in Fig. 45 is given.

H. J. Vander Bijl in his analysis[3] of this circuit gives the natural frequency of oscillation as

$$n = \frac{1}{2\pi}\sqrt{\frac{\left(1+\frac{r}{r_p}\right)}{L_2C}} \tag{14}$$

when r_p is the plate resistance and the remaining constants are given in the illustration.

Direct modulation by the usual method involves a cyclic change in the value of plate resistance. Hence, according to the above equation, there results a cyclic change in frequency which, though relatively small, becomes of the utmost importance when subjected to the peculiar phenomena of night-time transmission.

By making certain assumptions concerning the nature of frequency variation as amplitude modulation takes place, it is possible to work out distorted waves corresponding to various assumed wave interference conditions at the receiver. Perhaps the most simple and instructive means for producing these distorted waves is by a graphical method.

The equation for modulation of a high-frequency wave by a single tone may be written

$$e = (A + kA\cos vt)\sin pt \tag{15}$$

When A represents the unmodulated amplitude of the wave, k is a factor determined by degree of modulation, v is an angular velocity of the tone wave and p is the angular velocity of the high-frequency wave. The amplitude factor in this equation may be considered as a vector which is undergoing a change in length in accordance with the term included in the brackets. For the purpose of our analysis we shall include the angular velocity imparted to this vector by the last term in the above equation, since we are interested in the envelope of the resultant high-frequency wave at the receiver and the relative phase relations for two waves directly and indirectly transmitted combining to form this resultant. Since both carrier waves are of the same mean frequency only the relative position need be considered.

Now in our graphical determinations for the case of two transmission paths different in length, we represent the two effective fields by vectors varying in length in accordance with the amplitude factor of equation (15). However, due to the difference in length of path,

[3] "Thermionic Vacuum Tube," by Van der Bijl, page 274.

the changes in length of one vector will *lag the changes in length of the other* by an amount

$$\phi = v\,(\Delta t) \tag{16}$$

when Δt equals the difference in time of transmission over the two paths and v is the angular velocity of the modulating tone. This angle ϕ for 500-cycle modulation may according to the data thus far described, amount to more than 90 degrees at the receiving points selected for observation.

In addition to the lag in amplitude there will be a lag in frequency change over the frequency modulation cycle. This lag which has

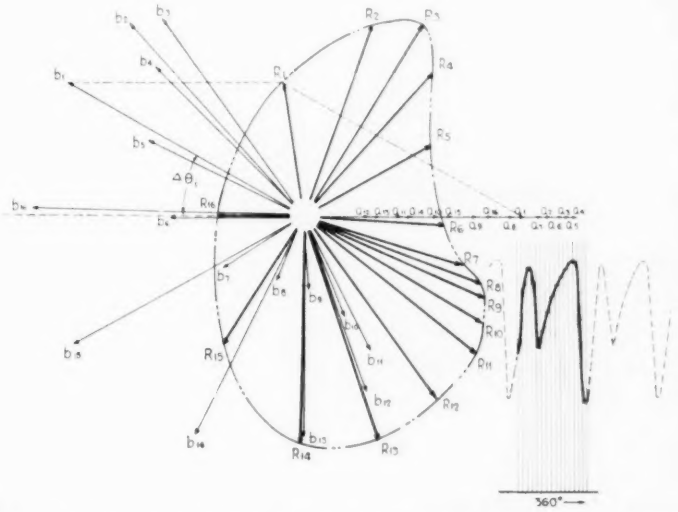


Fig. 46—Graphical method of synthesizing distorted wave forms caused by frequency modulation

already been shown in connection with the analysis of distortion in certain types of band fading records (see Fig. 22), becomes *a change in the relative phase angle* of the vectors under consideration. Thus our picture finally becomes one of two vectors changing in length, the changes in one continually lagging the changes in the other, the two vectors at the same time undergoing what we might term a relative angular wobble.

In Fig. 46 these relations are produced graphically. For our purposes we might assume that the vector representing one field is fixed and allow the other one to wobble the relative amount. At an

instant, for example, the directly transmitted field may be represented by a_1 in this figure. Assuming a difference in length of path, we may compute on the basis of the integral equation (13), the relative phase position of the vector representing the indirectly transmitted field b_1. The relative amplitude of this vector may also be determined by substituting $\Delta\phi$ in equation (15).

After establishing a sufficient number of vectors to represent the cyclic variation we may combine the respective components to obtain the resultant representative of the successive instants. These are

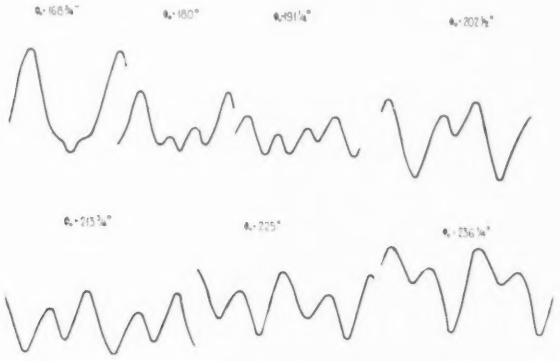


Fig. 47—Synthetic wave forms showing distortion due to frequency modulation

shown as R_1, R_2, R_3, etc., a broken line being drawn through their extremities to identify their positions. Now, if we plot these resultants as vertical ordinates in their successive time relation as shown on the lower right of Fig. 46, we have the envelope of the resultant wave at the receiver.

When the mean position of the two vectors (a) and (b) in Fig. 27 is 180 degrees separation, the signal is experiencing a fading minimum. When they are on the average in phase the amplitude is at a maximum. We can, therefore, trace a relation between quality distortion and fading by such an analysis, assuming a constant percentage modulation. Fig. 47 shows a series of high-frequency wave envelopes obtained by this method of graphic analysis. The mean vector relation is represented by ϕ_0, and for $\phi_0 = 180$ degrees the fading may be considered at a minimum. The waves shown in Fig. 47 being envelopes of the high frequency will undergo certain changes in the process of detection. These, however, would only slightly modify the wave.

For purposes of comparison, a set of oscillograph pictures of representative received wave shapes is shown in Fig. 48. These represent the actual effect of night-time transmission with frequency modulation between 463 West Street, New York City and Stamford, Conn.; the modulating tone was a practically pure 264-cycle sinusoidal wave. The samples have been arranged in successive order

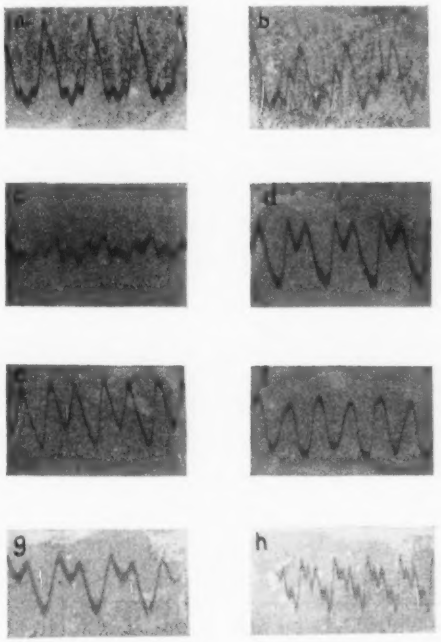


Fig. 48.—Oscillograms showing actual wave forms with distortion resulting from frequency modulation

to correspond with the order shown in Fig. 47. There exists a striking similarity. Occasionally, however, the shapes predicted may depart considerably from those obtained experimentally. As an example of such a departure, the record (h) in Fig. 48 has been included. Such unusual samples may be due to a combination of waves arriving over more than two paths or it may be that the time variation of the frequency is far from the simple sinusoid which we have assumed. As a matter of fact, a critical mathematical treatment of this case shows that only an approximation of such a sinusoidal condition is

possible since as has been shown by Carson,[4] a frequency modulated wave of this character consists of an infinite series of fixed frequencies spaced at regular intervals either side of a "fundamental" carrier wave. Obviously only a small part of such a series could get out of the transmitter or into the receiver due to circuit selectivity. For the lower modulating frequencies, however, the approximation involved in the assumption of a simple sinusoidal variation is not far wrong since the amplitudes of these side frequency components fall off rapidly as their order in the series increases. While 150 wave lengths difference in path length has been assumed for the synthesis of the wave shapes in Fig. 47, this difference may according to the data obtained amount to much more than this.

It may well be asked why this frequency modulation, since it produces such marked distortion at night in certain places, does not also give rise to distortion by day or in locations where transmission is steady. A full answer to this question would be far from simple. But in brief it is because the carrier and side-bands shift in absolute frequency together as a unit so that their relative or difference frequencies which determine the audio signal remain unchanged. Another way to put it is that the detector operates on the envelope of the high-frequency signals and is blind to the frequencies contained within the envelope except insofar as they affect the latter. However, since frequency modulation appreciably widens the frequency band occupied by the radio signals it is to be expected that the tuned circuits in the receiver would have some reaction on those louder portions of the signal for which the amplitude modulation and therefore the frequency modulation is large. The perfection with which broadcast signals may be received under suitable conditions leads one to believe that this effect must be small.

Fading in Relation to Form of Transmission

It has been shown that serious wave form distortion of the reproduced signal may result if frequency modulation occurs with the amplitude modulation and the transmission is subjected to night-time conditions. This distortion from frequency modulation can be eliminated by stabilizing the carrier frequency. There remain some wave form distortion and the annoying amplitude changes caused by selective fading which is one of the most serious present day problems in radio transmission. Let us now consider the nature and cause of this

[4] See "Notes on the Theory of Modulation," by John R. Carson, Proc. Institute of Radio Engineers, February, 1922.

residual wave form distortion and some further consequences of selective fading under the assumption that there is no frequency modulation involved.

The process of detecting audio signals from radio frequency signals is, at least in its simpler aspects, well understood, but it may not be generally appreciated that the action is such that the detected signals may be greatly modified by changes in the relative amplitudes and phases of the carrier and side-band components such as may result from their transmission through the medium. That the amplitudes and phases of the carrier and side-band signals are not necessarily received in the same relation that existed as they left the transmitter has been pointed out earlier, in the discussion on selective fading.

The usual expression for a high-frequency carrier wave of frequency $p/2\pi$ modulated by a low-frequency wave of frequency $v/2\pi$ is

$$e = A[1 + a \sin (vt + \phi)] \sin pt$$

where A is the carrier amplitude, a, the percentage modulation and ϕ the starting phase of the modulating tone with reference to the carrier. Expanded into its components this becomes

$$e = \frac{A_1 a}{2} \cos (pt + vt + \phi_1) \qquad \text{(the upper side band)}$$

$$- \frac{A_2 a}{2} \cos (pt - vt - \phi_2) \qquad \text{(the lower side band)}$$

$$+ A_3 \sin pt \qquad \text{(the carrier)}$$

where $\phi_1 = \phi_2 = \phi$ and $A_1 = A_2 = A_3 = A$ as the waves leave the transmitting antenna.

In the receiving set this function is squared by the action of the detector and, neglecting direct currents and frequencies above the audio range, the result is

$$\frac{a}{2} A_3 [A_1 \sin (vt + \phi_1) + A_2 \sin (vt + \phi_2)] + A_1 A_2 \frac{a^2}{4} \cos (2vt + \phi_1 + \phi_2) \quad (17)$$

of which the first term represents the fundamental frequency of the original modulating tone and the second term the second harmonic.

From this expression several conclusions can be immediately drawn. Due to the action of the detector there is always some slight wave form distortion as is evidenced by the presence in relatively small amplitude of the second harmonic. In the ordinary case this is negligible. The first term contains the carrier amplitude as a

factor but the second term does not. Thus, if selective fading erases the carrier at any time, reducing its amplitude to zero or a small value, the signal, represented by the fundamental tone, practically disappears, *even though the side-bands have not faded out*, and there remains only the harmonic. This is the residual distortion shown in Fig. 41 and which can often be heard during a fading out period. It is caused by the two side-bands beating together in the detector. We have here exposed a fundamental defect in the usual form of modulated signal transmission. The amplitude of the received signal is subject to all the whims of the carrier and to paraphrase freely an old saying we might remark that a signal is no stronger than its carrier. We may at once conclude that one way to reduce fading is to suppress the carrier and resupply a constant amplitude carrier at the receiving station.

Analyzing further the first term of the expression representing the detected signal, the first part of the bracketed portion results from beating together in the detector of the carrier and upper side-band and the second part from the carrier and lower side-band. It is clear that one of the side-bands may fade out completely and the other will still bring in the signal, provided the carrier is not also lost, with a phase shift to be sure but nevertheless not seriously reduced in amplitude. In telephony this kind of phase shift is relatively unimportant. Here we have an evident advantage in transmitting both side-bands since they support each other's frailties. But if the two side-bands suffer phase shifts in transmission, as we have earlier shown may be produced by wave interference, such that ϕ_1 and ϕ_2 differ by π radians or 180 degrees, the two components will cancel each other provided their amplitudes A_1 and A_2 remain equal. In other words all three components—carrier and both side-bands—may arrive at the receiver with full amplitude and yet no signal will be detected from them except a second harmonic component. This is obviously a disadvantage of transmitting both side-bands since, at such an instant, if one of them were eliminated the signal would reappear.

We conclude that there is, on the basis of such a brief analysis, not much to choose between single side-band and double side-band transmission when the carrier is transmitted also.

But if we wish to realize the advantages of carrier suppression a choice is not difficult. A carrier suppression system in which both side-bands are transmitted requires that the replacement of the carrier at the receiving station be done with almost absolute accuracy as to frequency and phase, a thing which involves very serious prac-

tical problems. On the other hand if but a single side-band is transmitted the difficulty is reduced to placing the carrier within a very few cycles of its correct position. The allowable departure will depend on a number of things but there is reason to believe that for high quality transmission it must be very small, perhaps no greater than two or three cycles.

With the single side-band carrier suppression method, invented by John R. Carson, the radiation is stripped down to the minimum which will fully transmit the telephonic signals and this reduces to a minimum the exposure of the signals to the ravages of selective fading. If the spacing interval of the fading is relatively narrow as in the cases we have examined hereinbefore, this form of transmission would not fade seriously in average volume but would be subjected to a continual changing of its frequency-amplitude characteristic, that is to say individual frequency components would fade progressively as the minima of the selective fading wandered back and forth across the frequency range encompassed by the single side-band. If the spacing interval of the fading were very large so that the minima were very broad of if some other, at present unexplored form of fading which covers a wide band at one time were acting, the signal would fade in average volume but the range of its variation would be only the square root of that of a carrier transmitted signal, since only the side-band would fade and the locally supplied carrier would remain unchanged.

The extent to which these theoretically drawn conclusions may be realized in practical application is yet to be determined but we have a few records bearing upon the matter which at least do not run contrary to them.

All of the transmission tests where the radio signal was beat with a local oscillator and the detected beat note observed, were equivalent to single frequency single side-band transmission with carrier suppression, the local oscillator functioning as the carrier suppressed at the transmitter. In this case, for which a number of records have already been shown, the detected signal is in proportion to the product of the amplitudes of beating oscillator and received radio signal. The phase of either does not affect the amplitude of the audio signal. Hence, the only important modification of the original signal is the variation in the amplitude resulting from selective fading.

Unfortunately we have no records in which a direct comparison is made between single side-band transmission with and without carrier suppression but the case can be visualized from the record shown in Fig. 12 or 13. Here each one of the frequencies recorded may be looked upon as a single side-band frequency which has been detected through

the agency of the resupplied carrier of the beating oscillator used to bring them down to audio-frequency. If now we were to take two of these frequencies shown on the record and multiply their amplitudes together at each point we would obtain the amplitude of the signal which would result if one of them were a single side-band and the other its accompanying carrier. It is obvious that the fading variations would thereby be increased in amplitude and rapidity.

In order to obtain a comprehensive picture of the relative advantages of radio transmission using a carrier and one side-band as compared

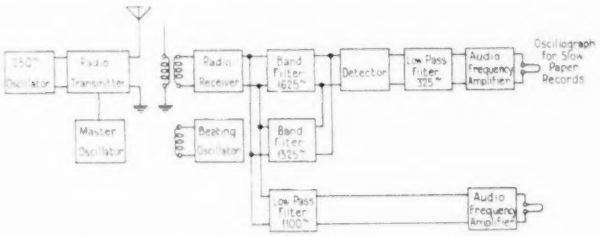


Fig. 49.—Diagram of system used to obtain records of transmission with carrier and one side band and carrier and both side-bands

with the common practice of transmitting both side-bands, the following tests were made. The schematic diagram of the circuit arrangement is shown in Fig. 49. At the transmitter the carrier and both side-bands are transmitted and at the receiver they were selected out by means of filters in the manner previously explained. The signals from the filters corresponding to the carrier and lower side-band were applied to the input of a detector circuit and from its output the detected difference signal was selected by a low-pass filter. This signal was equivalent to that which would be received if only the carrier and one side-band were transmitted. From the output of the radio receiver a branch circuit goes to a low-pass filter which transmits only the signal detected from the carrier and both side bands, suppressing from this circuit the higher frequency signals corresponding to carrier and side-bands produced by the beating oscillator and received signals.

By making simultaneously a record of these two signals a direct comparison is obtained of the effect of selective fading on their amplitudes. Fig. 50 shows samples of several such records made at Riverhead, L. I. The modulating frequency for strips 1, 2 and 3 is 250-cycles, and for strips 4 and 5, 500-cycles. The record on strip 3 is shown on account of the peculiar characteristic of the signal fading, for considerable periods of time remaining at relatively low amplitude.

In these oscillograms the upper trace is the record of the signal from the carrier and both side-bands, and the lower trace the signal from the carrier and lower side-band.

These records illustrate by giving a graphic comparison the effect of the phase changes of the component signals in the case where the

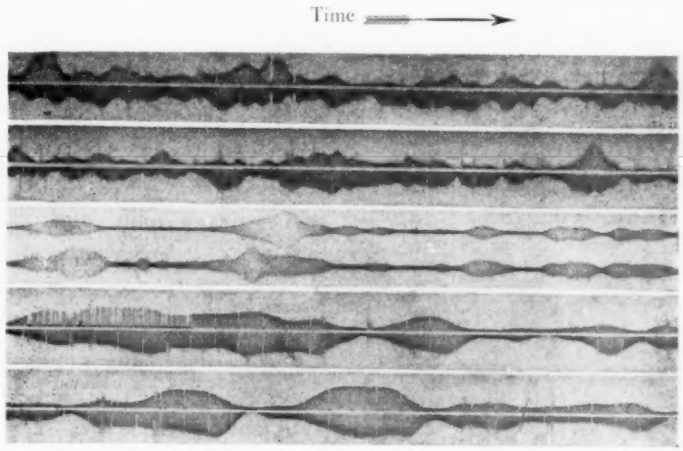


Fig. 50—Slow record comparing the signal detected from carrier and one side-band with signal detected from carrier and both side-bands. Made at Riverhead, L. I. Upper trace carrier + both side-bands, lower trace carrier + one side-band. Strips 1 and 2, July 22, 1925, 1:46 a.m. 250-cycle modulating tone. Strip 3, July 21, 1925, 3:10 a.m. 250-cycle modulating tone. Strips 4 and 5. July 23, 1925. 2:47 a.m., 500-cycle modulating tone

signal is detected from both side-bands. The amplitude of the signal from both side-bands in some instances is very small but appreciable amplitude is still indicated at the same instant for the signal from one side-band. This is explained as meaning that the side-band phases were such as to make the component signals 180 degrees out of phase after detection and that the amplitudes of the components were practically equal. The reverse situation is also observed where the amplitude of the signal detected from the lower side-band is zero and appreciable signal is recorded for the case where both side-bands are used. This is interpreted to mean that the side-band signal was eliminated by selective fading. In this event it was, of course, not contributing to the signal which was detected from both side-band signals. The recorded signal comes from the other side-band which evidently was not eliminated at that instant by selective fading.

Visual observations made with the cathode ray oscillograph, which unfortunately furnishes no permanent record of transient effects, confirmed the strip records in regard to the reality of there being side-band phase variations. From equation (17), it is seen that if these variations occur the fundamental of the detected tone signal at the receiver will not bear a fixed phase relation to that detected from the transmitting antenna current while if there are no such changes the phase between these two tones would remain constant. The locally detected tone and the tone detected from the transmitting antenna current and brought to the receiving station over telephone wires, were applied to the two pairs of deflecting plates in the cathode ray oscillograph Since the deflections caused by these two pairs of plates are at right angles to each other the resulting Lissajous figure from two sine waves of the same frequency will be a slanting line, an ellipse or a circle depending on their phase and amplitude relation. The actual figures were observed to change progressively through this range of shapes, the changes following roughly the magnitude and rapidity of the fading. The effect of amplitude changes on such figures is quite distinct from the effect of phase changes and there was no difficulty in separating out the evidence of large phase changes.

Considering only the above theories and facts there appears to be a reasonable basis for a conclusion that the best form of radio transmission for use in broadcasting is single side-band with carrier suppression. But on practical grounds we do not believe such a conclusion is justified. The fading and distortions which we have made much of in the preceding pages are not experienced by the majority of broadcast listeners when they listen to local stations. To require these listeners to provide themselves with more complicated and expensive receivers, simply to allow more distant or less favorably situated listeners to obtain better reception, seems neither reasonable nor desirable. The art offers several other possible avenues toward improvement much less difficult of application and it must be remembered that radio broadcasting is already reaching a degree of standardization and a volume of existing receiving equipment which rules that changes must come slowly and without serious prejudice to the existing order.

Conclusions

Subject to the limitations imposed by the scope of our investigations the following conclusions may be drawn:

Fading can be quite sharply selective as to frequency and the evidence points toward wave interference as the cause.

The evidence for wave interference indicates that some of the energy of received signals reaches its destinations by a circuitous route and suggests that this route is by way of upper atmospheric regions.

Quality distortion may result from dynamic instability of the transmitter.

Fixed wave interference patterns in connection with shadows sometimes exist in daytime transmission.

Abstracts of Bell System Technical Papers Not Appearing in this Journal

New Methods and Apparatus for Testing the Acuity of Hearing. HARVEY FLETCHER.[1] This paper presented before the American Otological Society, classifies hearing tests in four groups according to their purpose.

1. Industrial or those made for determining the fitness of a candidate for employment. In certain types of work it is particularly important that a prospective employee meet a definite requirement for acuity of hearing. Tests made in the army and navy for various branches of service are conspicuous examples of this kind of test.

2. Educational or those made for determining the degree of hearing of school children both in the public schools and in the schools for the deaf for the special purpose of determining the proper methods to be used in their education.

3. Clinical or those made for assisting the physician to make a proper diagnosis of the cause of deafness.

4. Research or those made to determine new facts about both normal and abnormal hearing.

It is highly desirable that a single scale be used for representing the degree of hearing which is independent of the method used and which has a general application to the four purposes enumerated. Such a scale is proposed and it is shown how the commonly made voice test, watch tick, acoumeter, coin click and tuning fork tests can be expressed in terms of hearing loss units on this scale.

The paper is concluded by summarizing the different methods for testing the acuity of hearing which are as follows: (1) voice tests, (2) phonograph audiometer, (3) hearing loss for speech calculated from audiogram, which audiogram may be obtained in three ways, (a) tuning forks (constant initial amplitude), (b) tuning forks (comparison with hearing of tester), (c) pitch range audiometer.

The Relation Between the Loudness of a Sound and Its Physical Stimulus. J. C. STEINBERG.[2] Experiments with many types of sounds have shown that the loudness of a sound is a function of its

[1] *The Laryngoscope*, Vol. XXXV, No. 7, July, 1925.
[2] *Physical Review*, Vol. 26, pp. 507–523, Oct., 1925.

energy frequency spectrum and its level above the threshold of hearing and that if this relationship be represented as

$$L = \frac{10}{3} \log_{10} \left[\sum_{i=1}^{i=k} (W_i P_i)^{2.r} \right]^{r}$$

sounds whose calculated values L are equal will appear equally loud to the average normal ear. P_i is the r.m.s. pressure of the ith component of the sound wave. The weight and root factors W and r, respectively, are functions of the sensation level, which is synonymous with the term loudness as formerly used and is defined as

$$S = 10 \log_{10} \left[\sum_{1}^{k} P_i^2 \Big/ \sum_{1}^{k} P_{oi}^2 \right]$$

where P_{oi} is the r.m.s. pressure of the ith component when the complex sound is at the threshold of hearing. In case the components in a narrow band of frequencies Δn are not resolved their energy must be integrated to obtain the energy of the equivalent single component. The root factor r is inversely proportional to the ratio of the minimum perceptible increase in energy to the total energy. For intensities near the threshold, the weight factors are equal to the reciprocals of the minimum audible pressures. Curves are given showing the values for W for various frequencies at various sensation levels, also the values of r as a function of S. As the intensity is increased the weight factors give greater weight to the lower frequencies; hence, even though the amplitude of the sound wave be increased without distortion, the ear will perceive both an increase and a distortion. This effect is due to the non-linearity of the ear.

Binaural Beats. C. E. LANE.[3] By introducing a tone of frequency f into one ear and another tone of frequency $f+N$ into the opposite ear, where N is less than 5 or 6 cycles, two kinds of binaural beats are obtained. Objective binaural beats are heard for most values of f within the audible frequency range, provided there is the proper difference in amplitude between the two tones. For telephone receivers as sound sources, this difference for best beats is about 55 TU and for the same receivers supplied with sponge-rubber cushions about 62 TU. These beats are heard because the louder tone is conducted through the head to the ear of the weaker tone and the two tones there are about equally loud. Subjective binaural beats are heard for frequencies below 800 or 1,000 cycles when the tones at the

[3] *Physical Review*, Vol. 26, No. 3, Sept., 1925.

two ears have about the same amplitudes, differing by not more than 25 TU. Data obtained with 22 observers are summarized. The evidence indicates that these beats are not due to cross conduction but are of central origin and the result of the sense of binaural localization of sound by phase. If the beats are slow (less than 1 per sec.) they are generally recognized as an alternate right and left localization, though some observers may report one or more intensity maxima during the beat cycle. Such maxima are explained as the result of one's interpreting the sound as louder when localization is more definite. Fast beats (more than 1 per sec.) are generally recognized as an intensity fluctuation. They are explained by assuming that the sound appears louder when the phase relations are such that it is normally best localized in the position toward which the attention is directed. This explanation is supported by observations made with a constant source rotating around the head of a listener.

Effect of Tension Upon Magnetization and Magnetic Hysteresis in Permalloy. O. E. Buckley and L. W. McKeehan.[4] Wires of five nickel-iron alloys containing 45, 65, 78.5, 81 and 84 per cent. Ni, 60 cm. long and 0.1 cm. in diameter, were studied by a ballistic method, for tensions up to 10,000 lb. per in.^{2} and fields up to saturation (10 to 20 gauss). Permalloy with 81 per cent. Ni is nearly indifferent to tension in its magnetic behavior; permalloy with less nickel is more easily magnetized and has less hysteresis when under tension, while 84 per cent. permalloy is more difficultly magnetized and has greater hysteresis when under tension. The saturation values are independent of the tension. In 78.5 per cent. permalloy, under a tension of 3,560 lb. per in.^{2}, saturation is reached at only 2 gauss (and is practically complete at 0.2 gauss) and the hysteresis loss is only 80 ergs per cm.^{3} per cycle, so small that it may be regarded as due to slight inhomogeneity rather than to any essential features of the magnetization process. *Relation to crystal orientation.* X-ray examination proves that this abnormally low loss is not due to any peculiar orientation of the crystal axes as the crystals are found to be oriented at random. Magnetostriction behavior can be deduced from these results. Above 81 per cent. Ni, permalloy contracts like Ni while below 81 per cent. Ni, permalloy expands like Fe.

Demagnetizing factor for a wire with a length 600 times the diameter, was determined experimentally and found to vary from a maximum of 1.6×10^{-4} to a low value, the changes being like these previously described by Benedicks for iron.

[4] *Physical Review*, Vol. 26, No. 2, Aug., 1925.

A Contribution to the Theory of Ferromagnetism. L. W. McKeehan.[5] *Relation of permeability and hysteresis to atomic magnetostriction.*—In permalloy, it has been found that magnetostriction changes sign at about 81 per cent. Ni, hysteresis losses can be made vanishingly small near this composition, and these effects are not due to the special alignment of crystals. It is suggested that in every ferromagnetic material the process of magnetization involves (1) intra-atomic changes, presumably changes in the orientation of electron orbits, governed by quantum dynamics and independent of environment; and (2) inter-atomic changes (stresses and strains). The interdependence of the inter-atomic changes and the intra-atomic changes is conveniently described as atomic magnetostriction. On this view, hysteresis loss and magnetic hardness are due to the energy required to produce, in succession, the local deformations associated with changes in the magnetization of single atoms or small groups of atoms. High initial permeability and low hysteresis loss in permalloy are explained as resulting from locally compensatory atomic magnetostrictions of the nickel and iron atoms in small groups. The fundamental differences in the magnetic behavior of Fe, Ni and Co are attributed to differences in their atomic magnetostrictions. Other differences are attributed to differences in the mechanical properties which alter the energy expended when atomic magnetostriction takes place.

Induction from Street Lighting Circuits: Effects on Telephone Circuits. R. G. McCurdy.[6] Synopsis.—This paper discusses series street lighting circuits from the point of view of their relations to nearby telephone circuits. These lighting circuits often have a much greater inductive influence in proportion to the amount of power transmitted than have most other types of power distribution or transmission circuits. This is due to the relatively large distortion in wave shape of voltage and current on certain types of these lighting circuits, and to the unbalanced voltages to ground which occur with series layouts. Three general types of lighting circuits are discussed. These are a-c, arc circuits, d-c, arc circuits supplied by mercury arc rectifiers, and alternating-current incandescent circuits. Of these, the incandescent type of circuit, in which the lamps are equipped with individual series transformers or auto-transformers, is the most important in this respect. Measures for reducing interference from these circuits are discussed.

[5] *Physical Review*, Vol. 26, No. 2, Aug., 1925.
[6] *A. I. E. E. Journal*, Vol. 44, pp. 1088–1094, Oct., 1925.

Power Distribution and Telephone Circuits. Inductive and Physical Relations. H. M. Trueblood and D. I. Coxe.[7] Consideration of the relation between power distribution and telephone systems is naturally involved in the comprehensive review of the problems of the rapidly expanding power distribution networks in this country. Pending the completion of studies now being actively carried on in this comprehensive review, a preliminary and qualitative discussion is given.

Situations of exposure fall into three groups determined by the character of the area served. (1) "downtown" districts; (2) residential urban districts; (3) rural districts. The major problems arise in the second group. A wide variety of arrangements characterize both systems, and require consideration.

Among technical features, coefficients of induction for close exposures, shielding action of metallic cable sheaths for both power and telephone circuits, and "ground potential" effects, are distinctive problems. Where both classes of circuits are in cable with suitable precautions as to grounding, interference is rarely to be anticipated.

Noise induction from power-distribution circuits is chiefly from residuals, which occur on single-phase branches of polyphase circuits, or where triple harmonics or load-current unbalances are introduced by grounding neutrals, or where admittances to ground of phase wires are unequal. Residual currents are largest in systems having multiple-grounded neutrals, both load currents and triple harmonics occurring. Approximate resonance at triple harmonic frequencies between the inductance of station apparatus and power cable capacitance has characterized several situations. Various single, two and three-phase arrangements are compared from the induction standpoint.

The closely related matter of unbalances in the telephone plant is briefly discussed.

[7] *Journal of the A. I. E. E.*, Vol. XLIV, No. 12, Dec., 1925.

Contributors to this Issue

BANCROFT GHERARDI, M.E., M.M.E., Cornell University. Engineering assistant, 1895–09; traffic engineer, 1899, New York Telephone Company; chief engineer, New York and New Jersey Telephone Company, 1900–06; assistant chief engineer, New York Telephone Company, and New York and New Jersey Telephone Company, 1906–07; equipment engineer, American Telephone and Telegraph Company, 1907–09; engineer of plant, 1909–18; acting chief engineer, 1918–19; chief engineer, 1919–20; vice-president and chief engineer, 1920—. Mr. Gherardi's work in the field of telephony is too well known to require comment.

ROBERT W. KING, A.B., Cornell University, 1912; Ph.D., 1915; assistant and instructor in physics, Cornell, 1913–17; Engineering Department of the Western Electric Company, 1917–20; Department of Development and Research, American Telephone and Telegraph Company, 1920–21; Information Department, 1921—.

WALTER A. SHEWHART, A.B., University of Illinois, 1913; A.M., 1914; Ph.D., University of California, 1917; Engineering Department, Western Electric Company, 1918–24; Bell Telephone Laboratories, Inc., 1925—. Mr. Shewhart has been engaged in the study of the relationship between the microphonic and physicochemical properties of carbon.

HARVEY FLETCHER, B.S., Brigham Young, 1907; Ph.D., Chicago, 1911; instructor of physics, Brigham Young, 1907–08; Chicago, 1909–10; Professor, Brigham Young, 1911–16; Engineering Department, Western Electric Company, 1916–24; Bell Telephone Laboratories, Inc., 1925—. During recent years, Dr. Fletcher has conducted extensive investigations in the fields of speech and audition.

JOHN R. CARSON, B.S., Princeton, 1907; E.E., 1909; M.S., 1912; Research Department, Westinghouse Electric and Manufacturing Company, 1910–12; instructor of physics and electrical engineering, Princeton, 1912–14; American Telephone and Telegraph Company, Engineering Department, 1914–15; Patent Department, 1916–17; Engineering Department, 1918; Department of Development and Research, 1919—. Mr. Carson's work has been along theoretical lines and he has published many papers on theory of electric circuits and electric wave propagation.

KARL K. DARROW, S.B., University of Chicago, 1911; University of Paris, 1911–12; University of Berlin, 1912; Ph.D., in physics and mathematics, University of Chicago, 1917; Engineering Department, Western Electric Company, 1917–24; Bell Telephone Laboratories, Inc., 1925—. Mr. Darrow has been engaged largely in preparing studies and analyses of published research in various fields of physics.

RALPH BOWN, M.E., 1913, M.M.E., 1915, Ph.D., 1917, Cornell University, Captain Signal Corps, U. S. Army, 1917–19; Department of Development and Research, American Telephone and Telegraph Company, 1919—. Mr. Bown has been in charge of work relating to radio transmission development problems.

DELOSS K. MARTIN, B.S., Polytechnical College of Engineering, 1920; U. S. Navy, 1918–1919; Department of Development and Research, American Telephone and Telegraph Company, 1919—. Mr. Martin's work has related particularly to radio broadcast transmission.

RALPH K. POTTER, B.S., Whitman College, 1917; E.E., Columbia University, 1923; U. S. Army, 1917–19; Department of Development and Research, American Telephone and Telegraph Company, 1923—. Mr. Potter has been engaged in experimental work relating to radio transmission phenomena.